BIOLOGICAL AND CLINICAL DATA INTEGRATION AND ITS APPLICATIONS IN HEALTHCARE

A Thesis
Presented to
The Academic Faculty

by

Matthew S. Hagen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing

Georgia Institute of Technology
December 2014

Copyright © 2014 by Matthew S. Hagen

BIOLOGICAL AND CLINICAL DATA INTEGRATION AND ITS APPLICATIONS IN HEALTHCARE

Approved by:

Professor Eva K. Lee,
Committee Chair
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor King Jordan
School of Biology
Georgia Institute of Technology

Professor Le Song
College of Computing
Georgia Institute of Technology

Professor Shamkant B. Navathe
College of Computing
Georgia Institute of Technology

Professor Timothy G. Buchman
School of Medicine
Emory University

Date Approved: 19 August 2014

To Sara, Mom, Dad and my sister,

it could not have been possible without you.

ACKNOWLEDGEMENTS

First and foremost, I would like to give the deepest gratitude and thanks to my advisor, Dr. Eva K. Lee. There were many moments during the course of my doctoral studies when I could not imagine myself sitting here writing the final lines of my dissertation. The largest lesson I learned from my advisor was not to worry or focus on a situation that may be causing stress. Instead, it was more important to keep calm and keep working. Initially entering the Bioinformatics PhD program at Georgia Institute of Technology, I wanted to learn how the knowledge I received from my Bachelors in Computer Science could be applied to the health sciences. There are many possible avenues that can be taken as a bioinformatics student, and I feel exceptionally fortunate to have the opportunity to work with Dr. Eva K. Lee in the exact subject area I was hoping to find when beginning my graduate studies. I am eternally grateful and hope to keep in close contact during the course of my career.

I also wanted to thank Dr. Timothy Buchman for serving on my committee and for the exposure to the Emory Healthcare department. These first experiences helped introduce me to the inner workings of hospital environments and the knowledge learned was invaluable. My Bachelors degree was at Georgia Institute of Technology as well, and my first database class was actually taught by my committee member Dr. Shamkant B. Navathe. I believe this was really where my excitement began for all the powerful features of relational databases. Perhaps, I just spent far too many hours beforehand trying to accomplish simple tasks with spreadsheets. Before taking Advanced Machine Learning with Dr. Le Song, I only really knew the basic algorithms without being fully aware of how much more could be done after learning more sophisticated methods. I wanted to thank Dr. Song for introducing me to diffusion kernels, as it made making predictions so much easier. Dr. King Jordan was one of the first professors I had met at the very beginning of my graduate studies. I loved the word bioinformatics and its definition, but I did not know exactly how the field transpired into the real world. The computational genomics class I took with Dr. King Jordan was exceptional, and it gave a great impression of how many different ways bioinformatics tools could be used to study diseases such as sequencing, gene prediction, and functional annotation. Most importantly, Dr. Jordan always has a way of making any topic fun and entertaining regardless of the subject.

Finally there's my family. First, I wanted to thank Sara. Thank you so much for your support and sticking this through with me the whole way. I can't say how great it is when you are feeling down and can't get your program working to have someone cheer you up with a home-cooked meal better than you would find at any restaurant. To my Mom, Dad, and sister, Vanessa, I do not think I would be here if not for you. I am indebited to my parents and sister for instilling a level of confidence in me that I do not believe I would have attained without them. Without their influence, I do not think I would have considered myself as someone capable to start a PhD program let alone finish one. Thank you for having no doubts. For anyone else I am forgetting, to my friends, my grandmother (Mum Mum), extended relatives, or anyone else, I wanted to thank you for not being angry for forgetting to mention you above. It is very late and its time to write the final period in this dissertation. All the best.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Answers to the most complex biological questions are rarely determined solely from the experimental evidence. It requires subsequent analysis of many data sources that are often heterogeneous. Most biological data repositories focus on providing only one particular type of data, such as sequences, molecular interactions, protein structure, or gene expression. In many cases, it is required for researchers to visit several different databases to answer one scientific question. It is essential to develop strategies to integrate disparate biological data sources that are efficient and seamless to facilitate the discovery of novel associations and validate existing hypotheses.

This thesis presents the design and development of different integration strategies of biological and clinical systems. The BioSPIDA system is a data warehousing solution that integrates many NCBI databases and other biological sources on protein sequences, protein domains, and biological pathways. It utilizes a universal parser facilitating integration without developing separate source code for each data site. This enables users to execute fine-grained queries that can filter genes by their protein interactions, gene expressions, functional annotation, and protein domain representation. Relational databases can powerfully return and generate quickly filtered results to research questions, but they are not the most suitable solution in all cases. Clinical patients and genes are typically annotated by concepts in hierarchical ontologies and performance of relational databases are weakened considerably when traversing and representing graph structures. This thesis illustrates when relational databases are most suitable as well as comparing the performance benchmarks of semantic web technologies and graph databases when comparing ontological concepts.

Several approaches of analyzing integrated data will be discussed to demonstrate the advantages over dependencies on remote data centers. Intensive Care Patients are prioritized by their length of stay and their severity class is estimated by their diagnosis to help minimize wait time and preferentially treat patients by their condition. In a separate study, semantic clustering of patients is conducted by integrating a clinical database and a medical ontology to help identify multi-morbidity patterns. In the biological area, gene pathways, protein interaction networks, and functional annotation are integrated to help predict and prioritize candidate disease genes. This thesis will present the results that were able to be generated from each project through utilizing a local repository of genes, functional annotations, protein interactions, clinical patients, and medical ontologies.

CHAPTER I

INTRODUCTION

1.1 Overview

As the volume of biological data sources continues to grow, there is a rising demand for an efficient system that can accurately integrate pertinent information. Researchers are commonly faced with challenges to answer complex biological questions, and the number of available databases has risen from 96 in 2001 to 1,552 in 2014 [147]. Specific studies can require joining many databases before an analysis can be conducted. Scientists may spend 80% of their time retrieving and pre-processing their data [414]. For example, verification of whether a particular gene sequence is associated with a specific biological process or structure would require visiting three separate data centers on gene sequence, protein sequence, and macromolecular structure. NCBI provides a comprehensive volume of database resources that include DNA sequences (Genbank) [40], annotated genes (Entrez Gene) [313], protein sequences (RefSeq) [398], protein domains (CDD) [317], 3-dimensional molecular structures (MMDB) [520], chemical molecules (PubChem) [521], and biomedical literature (Pubmed).

It is not a trivial task to centralize data from a variety of sources into one location with uniform access. Wang et al. cites data integration as one of the most important research areas, but it is faced with the challenges of heterogeneity, autonomy and maintainability [522]. Separate resource sites uses their own structure to store and represent their biological data. It is not possible to enforce a standard model that is used across all possible biological data types. The environment is constantly changing and as new discoveries are made, new data structures are created to meet shifting demands.

In navigational approaches, users must manually join two data sources with a “point and click” interface. It is not realistic to ask scientists to tediously inspect all returned records for a given query [182]. Data warehouses attempt to alleviate these demands by integrating all relevant information into a local repository. All queries are executed on a local machine rather than an external server. Service-based approaches do not require investigators to store any data and provides a virtual view of the integrated sources. There is an ongoing debate as to which approach is most appropriate to handle the data integration requirements of a complex environment that is constantly evolving. Service-based approaches are well suited for cases where researchers need to integrate and search many data sources that are all up-to-date and undergoing rapid changes [304]. Data warehouses require serious maintenance challenges to regularly match data from external sources that are constantly updated. However, user queries yield excellent response times and facilitate fine granular access to the data [453]. Biological sciences contain many diverse data types and it is difficult to create a global schema that precisely captures all necessary features. Often, a great deal of richness in the data structures is lost in conversion to a relational model. Data warehouses can create a powerful system that delivers effective results, but they require extensive knowledge in database development as well as considerable setup-time to import each data source.

This dissertation attempts to alleviate the shortcomings of data warehousing while simultaneously taking advantage of the many compelling features that locally integrated databases bring the biological and clinical sciences. A universal parser automatically translates different biological resources without custom source code for each data model. Standardized hierarchical ontologies are integrated by utilizing the state of the art technologies in natural language processing and Graph databases. Applications of unified resources are presented in the biological domain by using diverse datasets to prioritize disease candidate genes. In the clinical domain, integration of patient databases with standardized ontologies facilitate improved prioritization of intensive care admissions and accurate clustering of multimorbidity conditions.

1.2 Contributions of this thesis

The work presented in this thesis represent a number of contributions to the data integration and analysis techniques for the biological and clinical sciences:

- A relational translator (BioSPIDA) can convert public biological databases in ASN.1 or XML format to local relational databases without custom source code as long as a schema file is provided for each resource. The implementation is comparatively lightweight to other data warehousing solutions, because of parallelization, parsing with efficient memory management, and removal of redundant data.

- Disease candidate genes can be prioritized by many additional data sources, due to integrated biological data. A local warehouse contains information on pathways, literature, gene expression, protein domains, and gene ontology.

- Techniques are introduced to improve healthcare with the data analysis of Electronic Health Records and the integration of standardized medical terminologies. This includes the prioritization of intensive care admissions and the clustering of multimorbidity conditions.

1.3 Structure of this thesis

Chapter 2 This chapter discusses a background of the many technologies utilized in this dissertation. Data centers can vary in their integration methods as well in the structures that are used to represent records. Many distance metrics are used in this thesis, and there is a large variety of semantic similarity measures for concept graphs. The current approaches for data representation, data integration and semantic similarity measures are reviewed.

Chapter 3 This chapter covers the BioSPIDA system and the development of the final integrated data warehouse of many separate biological resources. BioSPIDA was used to convert NCBI databases, and additional databases were constructed from other external resources. The details of how XML and ASN.1 records are efficiently translated into a relational database are discussed, and the advantages of rapid query execution speeds are illustrated.

Chapter 4 This chapter introduces a method to prioritize candidate disease genes by the integration of many data sources. The biological sources are provided by the data warehouse built in Chapter 3. A diffusion kernel matrix is constructed for five separate data types on pathways, literature, gene expression, protein domains, and gene ontology. The similarities between genes for each of these data types are calculated from different measures ranging from semantic similarity to vector cosine distance.

Chapter 5 This chapter discusses a clinical integration project that utilizes publicly accessible patient databases and standardized medical terminologies. Natural language processing is applied to diagnosis records to map terms to ontological concepts. A Graph database is built to fully convert a medical ontology for semantic similarity measures to accurately find commonalities between patients.

Chapter 6 This chapter develops a prioritization method for intensive care patients by the natural language processing of diagnosis records with linking to terms from a medical ontology. Patient diagnosis concepts can be recognized, and their length of stay can be predicted by assessing their severity level. Patients in different categories of length of stay and severity level can be accurately prioritized under different ordering schemes.

Chapter 7 This chapter introduces a density-based clustering method that can effectively place multimorbidity patients into similarly related groups. With the usage of Graph Databases, semantic similarity measures facilitate robust distance metrics that can accurately compute the strength of common patient characteristics.

Chapter 8 This chapter is the conclusion of the thesis where the current work is summarized and potential avenues for future work are introduced.

CHAPTER II

BACKGROUND

Several major approaches of database integration exist in the bioinformatics domain, but there is no single best approach that outperforms all others. Each method has its own challenges, and it is important to review the existing data integration regimes to highlight the advantages and disadvantages of each strategy. Data sources can also be represented with a variety of different models that have their own strengths and weaknesses. This chapter discusses the different data integration and data representation regimes with providing examples of successful research projects.

Semantic similarity measures are another important topic in the biomedical domain. Terms in large volumes of unstructured textual resources can be mapped to standardized concepts. Semantic similarity measures can be used to calculate the degree of relatedness between these terms. However, the accuracy of these measures can depend on the corpus coverage as well as the specific domain. The last section of this chapter covers many of the state of art approaches for calculating semantic similarity.

2.1 Data Representation

Data integration is mainly concerned where data resides and how the different data sources are interconnected with each other, but there is a great deal of variation on how the data and knowledge may be represented at each of these data sources. The first section in this chapter will discuss models for data represention and their disadvantages and advantages.

2.1.1 Relational schemas

The relational model [102] of a database is organized into terms of tuples (rows), attributes (columns) and relations (tables). Relational data models are a mature and battle tested technology and are well understood by many industry professionals. However, its technology is not without criticisms especially for the biological data sciences. Biological entities can be complex hierarchical objects and modelling into relational objects is not trivial even for the most experienced database designers [277]. Objects must be explicitly structured and precise, and in the biological data sciences, relationships are usually ambiguous and not completely understood [304].

2.1.2 Semi-structured data

Semi-structured data frees data modelling from the rigid constraints of relational databases. It can be viewed as a series of labels with associated values, and it supports the nesting of labels. XML is a format supported by the WWW Consortium [56] and is an example of semi-structured data. XML allows for a more natural modelling of biological entities, because it highly flexible and has rich capabilities for linking data [1]. However, it cannot model complex relationships, such as many-to-many links between entities which are required to accurately describe pathways [304]. The Semantic Web is a data model that can use XML as its syntax and will be discussed in Section 2.2.6. The Semantic Web can be viewed as a paradigm shift where the World Wide Web is interconnected by pages in a machine-interpretable format [43], rather than a set of documents solely for keyword search and retrieval. Each page is expressed in RDF [337], which is a semi-structured data model that can describe arbitrarily complex relationships.

2.1.3 Ontologies

Ontologies are an explicit specification of concepts and their representations for a particular field of study [195]. The World Wide Web Consortium (W3C) provides several data models, including Resource Description Framework (RDF) [337], RDF schema (RDFS) [320], Web Ontology Language (OWL) [328], and the standard web query language for RDF (SPARQL) [379]. RDF is described as a series of triples, containing a subject, object, and predicate. The Semantic Web expresses ontologies in OWL format [328] which is built on top of RDF. An ontology is a set of concepts within a domain with a shared vocabulary that expresses the properties and relationships of each concept [195]. Ontologies are greatly beneficial in integrating data that span multiple disciplines, because they help to resolve semantic inconsistencies which are especially evident in the biomedical community [304].

Two regularly used ontologies in the biomedical community are Gene Ontology (GO) [20] and SNOMED CT [472]. The Gene Ontology (GO) consortium describe biological processes, cellular components, and molecular functions with the goal of producing a gene product vocabulary for all organisms. Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) is an ontological resource for the clinical domain that is used for clinical decision support, ICU monitoring, indexing medical records, medical research, and disease surveillance. By expressing the relationships among the various concepts, the ontology organizes the concepts in a structured way, thus reducing the number of possible interpretations of these concepts.

2.2 Data Integration

This section will review the most common data integration approaches in the biological sciences and also give examples of present approaches for each methodology. In some cases, it can be preferable to query external sources for information, while in others it is beneficial to build a locally accessible repository. The strengths and weaknesses of each strategy as well as the appropriate development settings will be discussed.

2.2.1 Navigational Integration

Navigational integration arose from the fact that an increasing number of sources on the web require users to visit and browse through several pages and hyperlinks before they reach their desired information [114]. Its justification is that many of these sources would not be accessible or found without point-and-click navigation [161]. Its goal is to facilitate the discovery of deeply linked pages by performing a quick search of related links. Sources are defined as sets of pages, and queries detect all possible paths between entities [60,161,345]. Relational databases do not permit the answering of such queries, but navigational integration is not a true integration of sources. It is particularly vulnerable to naming ambiguities which are frequent in the biomedical community. It also requires cooperation between all of the data centers [315].

The Sequence Retrieval System (SRS) [138] parses flat files that contain structured text with field names. It keeps track of cross-references between sources by storing local indexes for each field. After retrieving relevant entries at query-time, these links are utilized to suggest more results to the user. The Entrez system [369] is an access point to all sources provided by the National Center for Biotechnology Information (NCBI)[1]. Similar to SRS, different data sources are linked, so returned entries also have related links to other sources. BioNavigator[2] is a commercial solution that allows users to specify their preferred execution path when submitting queries. These execution paths can be reused later for similar type queries.

2.2.2 Data Warehousing

Data warehousing consolidates all of the designated data for research projects into one consolidated location. The data is migrated from multiple data sources into the central repository. The imported data can be cleaned, filtered, or transformed

[1] http://www.ncbi.nlm.nih.gov/

[2] http://www.bionavigator.com/

to meet the specific needs of the project, and it is then loaded into the warehouse. Executed queries are not evaluated at the sources, but on the local schema. This creates fast access and excellent response time to user queries. This is not trivial, because performance is often cited as a primary feature by biologists [199]. Since all of the data is locally housed, it avoids problems of low network response time and unavailability of sources. There are also very important benefits to research projects that have the ability to validate and transform data prior to integration [115].

Despite these advantages, data warehousing has its shortcomings. The warehouse must be regularly reloaded or it runs the risk of returning outdated results. Changes in the data model of the sources requires continuous modifications of the local schema. Many biological data sources change their data structure roughly twice a year [473]. The volume of the data may also be too large for the warehouse to handle. Only the most experienced may be capable to create a global schema that captures the data models of all the desired sources, and the richness of the individual data sources can be lost [304]. Warehouses may be best suited to projects that are focused on a very specific and narrow area of research.

IGD [417] integrated over 20 different data sources into one huge relational database. Since it had its own global schema, specific parsers had to be written to translate the schema for each of the individual sources. It was difficult to maintain the system and was discontinued. Atlas [453] integrates local instances of biological sequences, ontologies, and molecular interactions including: GenBank [41], RefSeq [400], UniProt [28], Human Protein Reference Database (HPRD) [397], Biomolecular Interaction Network Database (BIND) [27], Database of INteracting Proteins (DIP) [536], Molecular Interactions Database (MINT) [76], IntAct [258], NCBI Taxonomy, Gene Ontology (GO) [20], Online Mendelian Inheritance in Man (OMIM) [206], LocusLink [399], Entrez Gene [314], and HomoloGene [314]. It stores similar types using common data models and utilizes a C++ API to perform sequence and feature retrieval tasks. With integrated information, new associations could be derived, such as projecting known protein-protein interactions in one species onto another. Other examples of successful biological data warehouse projects are University of California at Santa Cruz (UCSC) Genome Browser [253], Genomics Unified Schema (GUS) [113], BIOZON [45], and MoDA [354]. BioWarehouse [288] and BioDWH [494] are examples of open source toolkits for locally constructing data warehouses from biological data sources.

2.2.3 Mediated Integration

Mediated integration focuses on query translation, as opposed to data warehouses that are focused on data translation. All of the data is left at the source. Queries are translated into sub-queries against many data sources. Query translation systems retrieve the returned data from the disparate data sources and abstractly present the information to the end user. They provide immediate access to up-to-date data from multiple data sources [546], but data cleansing can be difficult because it is not housed locally [347]. They are relatively inexpensive, because the data resides at its sources, but rely heavily on network connectivity and must be maintained when changes are made by the data sources.

According to surveys of database integration in biomedicine [10,479], there are four main classes of query translation systems that manage integration between the data sources: i.) pure mediation, ii.) global conceptual schema, iii.) multiple conceptual schemas, iv.) hybrid approaches. Global-as-view (GAV) [499] and Local-as-view (LAV) [292] are also used in the literature. GAV is equivalent to global conceptual schema and LAV matches with multiple conceptual schemas.

2.2.3.1 Pure Mediation

Pure mediation, also referred as Federated Databases [198], use developed software for each integrated database to accurately transform the data from the sources. The software, also known as "wrappers," must be carefully constructed according to the domain and features of the underlying data sources. In pure mediation, any data source change requires revisions and recompilation of the wrapper source code. Federated databases can be seen as a middle-ground between no integration and total integration and came before more advanced methods of mediator-based integration. They use a common data model, so it is difficult to represent diverse data types that represent all data sources. Federated Databases are most appropriate for research that requires the most up-to-date information from many related public data sources.

BioKleisli [96] was a pioneer in applying the federated database approach to the biomedical sciences. It was able to answer queries, such as finding many non-human homologs of genes located on particular human chromosomes. It uses a high-level query language, Collection Programming Language (CPL), that is more expressive than SQL and can query against several sources. The data model is object-oriented and is more expressive than a relational model, including lists, variants and nested records. It is well suited to handle complex data types common in biological entities. However, CPL requires source specific wrappers for each integrated database, and there is no specific molecular biology ontology to help formulate queries. This requires extensive knowledge of the structure of the data sources.

BioMart is a query-oriented data management system where each participating database in the federation must import their data into a local BioMart implementation. AceDB was a federated database that provided a common schema for multiple small genome projects. Other examples of federated databases include OPM [82] and GMOD [474].

2.2.3.2 Global Conceptual Schema

Federated databases can be challenging because there are many source schemas of the different data sources used in the federation. Mediator-based integration address this problem by utilizing a mediator schema that describes the sources and the relations among them [315, 523]. A mediated schema is generally a graph with entities represented as nodes and relationships between entities represented as edges. Queries are acted directly against the mediated schema and not the union of the data sources. The mediator is responsible to reformulate the query on the corresponding data sources. This enables the user to ask general questions without knowing the details of the underlying data sources. Mediated schemas are best suited for researchers who benefit from the simpler abstraction and have to ask very complex questions that span many different data sources.

In a mediated global conceptual schema, also known as global as view (GAV), the information required to transform the data sources is stored separately from the source code of the wrapper. GAV stores a global schema that contains a common vocabulary with mappings to each data source. It is very easy to reformulate queries because they immediately unfold into the equivalent definitions of the data sources. However, it is much more difficult to add or remove a data source, because it involves a direct modification of the mediated schema. It can be problematic to conceptualize a mediated global conceptual schema, because sources can have large variation in their data models even if they are in the same domain.

DiscoveryLink [199], developed by IBM, is a mediated system that serves as an intermediary to several biological sources. It submits SQL queries to a global schema with a structured object-relational model. TAMBIS [29] is a mediator-based and ontology-driven system. It was built on top of the BioKleisli system. Queries are constructed from a global schema and then formulated into CPL. Unlike many other projects that utilize ontologies, the ontologies in TAMBIS are not used for schema mapping between the data sources. Instead, they define the relationships between concepts. Similar to TAMBIS, K2/Kleisli [113] adds a layer on top of Kleisli combining the component schemas into a global schema and making the source databases transparent to the user. The data model for TAMBIS is structured and object-relational.

K2/BioKleisli has a semi-structured object-oriented data model. Other mediated systems with global conceptual schemas are ARIADNE [8], BioDataServer [158], BACIIS [336] and BIS [278].

2.2.3.3 Multiple Conceptual Schemas

In multiple conceptual schemas, also known as local as view (LAV), each source describes its own conceptual schema and are not required to share the same vocabulary. It is up to the data sources to describe the relations of the local schema in terms of the global mediated schema. The relations in the source schema are expressed in terms of a query in the mediated schema. This makes it very simple to add or remove sources, but query formulation is more complicated and intensive. The inter-schema links define semantic relationships among concepts, and it can be difficult to model in highly heterogeneous cases of the data sources. LAV is the preferred method for large-scale integration because any changes of the data sources have a lower impact. GAV is more suitable when the integrated data sources are well known and stable [214].

INDUS [63] integrates heterogeneous biological data sources utilizing ontologies as a semantic framework of the integrated data. Users may create their own ontologies to view a collection of data sources with tables structured according to their definitions. User queries can be answered without the need for a common global ontology. BioMediator uses a mediated schema to query across multiple structured and semi-structured data sources. Each user can create custom mediated schemata and ask diverse questions of the data sources. KIND [196] is a semi-structured object-oriented data model that builds an F-logic schema of the sources. Links between the integrated sources are governed by F-logic rules. In OntoFusion [382], virtual schemas are developed that represent ontologies of the data structure of the sources. The mapping process merges the virtual schemas into a unified global schema.

2.2.3.4 Hybrid Approaches

A hybrid approach is a compromise between global conceptual schemas and inter-schema links. Multiple conceptual schemas are still created, but a shared vocabulary is utilized as in a global conceptual schema when developing semantic specifications for each of the target sources. This hybrid approach simplifies the query translation process.

DataFoundry provides federated access to its data sources but caches the most frequently accessed records and maintains a global relational schema that can be queried with SQL [110]. TINet is a hybrid approach where most of its data sources are accessible by a federated model, except GenBank [41] and SwissProt [49] are maintained in centralized repositories [131]. SEMEDA has a relational database backend to store ontologies and database metadata while access to heterogeneous data sources is provided with federated systems [270].

2.2.4 Peer Data Management Systems

Peer Data Management Systems (PDMS) is viewed as the next evolutionary step in data integration systems. [193] As the number of data sources increases, it is increasingly difficult to develop a global mediated schema that encompasses the entire domain. PDMS decentralize sharing by allowing multiple mediated schemas that are integrated into the system. Peers provide data and a semantic mapping between one or a few other data sources. The PDMS can then traverse a semantic network to answer specific queries. They do not require stable data sources or a global mediated schema. Groups of users can create their own mediated schemas and appropriately map them to other peers on the network. They are the next step in mediated integration, but few biological projects have yet adapted to a peer data management system [304]. There are several competing projects in peer data management systems, including the Hyperion Project [15] and the Piazza project [204]. BioFuice [261] is a data integration project in bioinformatics which uses bidirectional peer mappings between sources. It can execute queries and aggregate information from sources with explicit links between peers.

2.2.5 Service-oriented Integration

Service-oriented approaches allow access to multiple heterogeneous sources through computer interoperability. Individual data sources grant access to their data via Web Services (WS). Contrary to Data Warehouses and Mediated Integration, Service-oriented Integration is a decentralized approach that supports interaction through a web application programming interface (API) [455]. They are several different Web Service Description Languages (WSDL). These include Simple Object Access Protocol (SOAP), and REpresentational State Transfer (REST). SOAP is a protocol that transfers XML-based messages over networks [54]. REST is a protocol that is implemented with HTTP methods [148]. Service-oriented approaches are advantageous to the field of bioinformatics, because they provide up-to-date information from various data sources [546]. However, it requires that many data sources provide WS as well as standardization in nomenclature and protocols. There are also the additional challenges of network connectivity and bottlenecks.

BioMoby [524] is an open source ontology-based WS for accessing heterogeneous data sources. It does not follow the RDF or OWL standards provided by W3C. This is in part due to the project originating in 2001 before these standards were stable. The MOBY Object Ontology contains IS-A, HAS-A, and HAS relationships between data formats. It utilizes SOAP for data exchange. Distributed Annotation System (DAS) [128] provides access to multiple distributed data sources using SOAP-based WS, and it is widely used in the genome annotation community. Taverna [223,362], a part of MyGrid [475], is a graphical workflow application that allows users to integrate many different software components, including SOAP or REST Web services. Taverna is aimed at integrating a growing number of biological tools and databases, and it has the ability to monitor running processes and data production. BioPython [75, 101] is a suite of python libraries that can carry out a wide range of bioinformatics problems. It can access key online databases, read sequence file formats and process 3D macro molecular structures.

2.2.6 Semantic Integration

In the biological community, most web pages are in a format easily readable by human beings (e.g. HTML). The goal of the Semantic Web is to describe information that can be understood by computer algorithms. Its aim is to create an interconnected network that can be easily processed, because of universal standards for data description [14]. Semantic Integration is an exciting addition to bioinformatics facilitating automatic machine reasoning and inferences with ontology-based integration [358]. However, they have similar limitations to Data Warehousing. It translates data from multiple sources all into a local repository in RDF format. It requires experienced database designers to convert the data sources into triples, and the translators must be periodically updated in the event of a data model change. The local warehouse must also be regularly updated to keep the data current. Semantic integration also does not have the speed advantages of a data warehouse. The data is densely stored as triples and it can not take advantage of the advanced indexing that a relational database would provide.

Bio2RDF [39] converts data into RDF format by applying Semantic Web technologies to various biological sources, such as Entrez Gene [314], KEGG [248], OMIM [206], PDB [42], and UniProt [28]. The RDF documents are linked with normalized Uniform Resource Identifiers (URI) and can be queried with SPARQL. YeastHub [92] integrates different types of yeast data in RDF format from data sources, such as SGD [90], MIPS [333], BIND [27] and GO [20]. Health Care and Life Sciences Interest

Group (HCLS) [93] aims to discover the benefits of the Semantic Web for the life sciences and translational research. It imports data from general sources, such as Entrez Gene [314], GO [20], HomoloGene[3], and also domain-specific sources such as the Allen Brain Atlas (ABA) [244]. Other applications of Semantic Web technologies in the Life Sciences are Cell Cycle Ontology (CCO) [13], BioGateway [12], CardioSHARE [511], NeuroCommons [425], Linked Life Data project [145], and GenoQuery [290].

2.3 Semantic Similarity

The final section in this background discusses the many different types of semantic similarity measures. When concepts are annotated to a standardized ontology, the difficulties of integrating heteregenous information can be alleviated by disambiguating biomedical terms. Semantic similarity measures are used to calculate the degree of relatedness between these terms, but there is no clear choice for the best strategy. The performance of these measures can depend on the chosen corpus, field of study, or evaluation benchmark. Measures can be calculated by edge counting, overlapping features, information content or hybrid methods, and this section reviews many of the state of the art approaches.

2.3.1 Edge based

Ontologies are a directed graph where nodes are connected by taxonomic (is-a) links, and in some cases non-taxonomic links. Rada et al [403] defines the distance between two concepts as the shortest path that links both entities in the ontology by is-a links:

$$dis_{rad}(c_1, c_2) = min_{\forall i} \left|path_i(c_1, c_2)\right| \tag{2.1}$$

where $|path(a, b)|$ is the length of the path that connects terms a and b. Pedersen et al [376] converts distance to similarity by inverting the path length between the

[3] http://www.ncbi.nlm.nih.gov/homologene/

concepts:

$$sim_{path}(c_1, c_2) = \frac{1}{dis_{rad}(c_1, c_2)} \tag{2.2}$$

Subsequently, various modifications and improvements have been made to edge-counting measures. Leacock & Chodorow (1998) (LCH) [282] define similarity as the ratio of path length to the maximum depth in the taxonomy with a logarithmic scaling:

$$sim_{lch}(c_1, c_2) = -log\left(\frac{\delta(c_1, c_2)}{2 \cdot max\_depth}\right) \tag{2.3}$$

where $\delta(c_1, c_2) = min_{\forall i} \, |path_i(c_1, c_2)|$, and max_depth is equal to the maximum depth in the taxonomy. The LCH similarity can also be rescaled to the unit interval by dividing by $log(2 \cdot max\_depth)$ [19]:

$$sim_{lch}^{scaled}(c_1, c_2) = 1 - \frac{log(\delta(c_1, c_2))}{log(2 \times max\_depth)} \tag{2.4}$$

Wu and Palmer (1994) [533] take into account the relative depth in the taxonomy of the concepts being evaluated. Deeper linked concepts are viewed as more similar, because specializations are more distinct. The relative depth of both concepts is represented by the depth of the least common subsumer $lcs(c_1, c_2)$, their closest common parent with maximum depth. Wu and Palmer define similarity with the following equation:

$$sim_{wp}(c_1, c_2) = \frac{2 \times depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \tag{2.5}$$

Pekar and Staab (2003) [377] calculate similarity as a ratio of the shortest paths from the terms and their LCS, and the shortest path from the LCS and the hierarchy root:

$$sim_{pk} = \frac{\delta(LCS(c_1, c_2), root)}{\delta(c_1, LCS(c_1, c_2)) + \delta(c_2, LCS(c_1, c_2)) + \delta(LCS(c_1, c_2), root)} \tag{2.6}$$

where $\delta(c_1, c_2)$ is the length of the number of edges of the shortest distance between c_1 and c_2.

2.3.2 Hybrid approaches

Hybrid approaches use path length in addition to other structural characteristics, such as local densities and relative depth of concepts. Li, Bander, and McLean (2003) [294] measure semantic similarity by combining shortest path length and the minimum depth of the LCS into a non-linear function:

$$sim_{li}(c_1, c_2) = e^{-\alpha \cdot \delta(c_1,c_2)} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \tag{2.7}$$

where $\delta(c_1, c_2)$ is equal to the minimum path length between c_1 and c_2, and h is equal to the minimum depth of their LCS. $\alpha \geq 0$ and $\beta \geq 0$ are parameters that scale the contribution of p and h. Al-Mubaid and Nguyen (2006) [6] also combine the minimum path-length with depth. A cluster is defined with each of the branches in the hierarchy. They introduce a common specificity C_{spec} measure (C_{spec}) that is the difference between the depth of a cluster D_c and the depth of the LCS:

$$C_{spec}(c_1, c_2) = D_c - depth(lcs(c_1, c_2)) \tag{2.8}$$

Al-Mubaid and Nguyen utilize the C_{spec} measure to weight lower level pairs of concepts as more similar than higher level pairs:

$$dis_{AMN}(c_1, c_2) = log((\delta(c_1, c_2) - 1)^{\alpha} \times (C_{spec})^{\beta} + k) \tag{2.9}$$

where α and β are tuning parameters and k is a constant. The authors use $\alpha = \beta = k = 1$. The hybrid approaches do provide higher accuracy as compared to pure path-based methods, but it can be rigorous to the tune the weights appropriately according to the input ontology [434].

2.3.3 Feature-based measures

Edge-counting measures treat taxonomical links with uniform distance. Feature-based methods attempt to overcome this limitation by measuring the degree of overlapping features between ontological concepts. These strategies are developed from the Tversky [498] model. The features of a concept are usually considered as the set of its subsumers including itself [207]. Let $\psi(c_1)$ and $\psi(c_2)$ represent the features of concept c_1 and c_2 respectively. Let $\Psi(c_1) \cap \Psi(c_2)$ be the shared features of concepts c_1 and c_2. Let $\Psi(c_1) \setminus \Psi(c_2)$ be the set obtained when eliminating features of c_2 from the features of c_1. Let$\Psi(c_2) \setminus \Psi(c_1)$ be the set obtained when eliminating features of c_1 from the features of c_2. The Tversky model defines similarity between c_1 and c_2 as the difference between shared and unshared features:

$$sim_{tve}(c_1, c_2) = \alpha \cdot F(\psi(c_1) \cap \psi(c_2)) - \beta \cdot F(\psi(c_1) \setminus \psi(c_2)) - \gamma \cdot F(\psi(c_2) \setminus \psi(c_1)) \quad (2.10)$$

where α, β, and γ are tuning parameters that weight each component. Batet [434] define taxonomic distance as the ratio between distinct and shared features:

$$dist_{Batet}(c_1, c_2) = log_2 \left(1 + \frac{|\Psi(c_1) \setminus \Psi(c_2)| + |\Psi(c_2) \setminus \Psi(c_1)|}{|\Psi(c_1) \setminus \Psi(c_2)| + |\Psi(c_2) \setminus \Psi(c_1)| + |\Psi(c_1) \cap \Psi(c_2)|} \right) \quad (2.11)$$

In Rodriguez and Egenhofer [420], the semantic similarity is computed by:

$$sim_{RE}(c_1, c_2) = \frac{|\Psi(c_1) \cap \Psi(c_2)|}{\gamma \cdot |\Psi(c_1) \setminus \Psi(c_2)| + (1 - \gamma) \cdot |\Psi(c_2) \setminus \Psi(c_1)| + |\Psi(c_1) \cap \Psi(c_2)|} \quad (2.12)$$

where $\gamma \epsilon [0, 1]$ is a parameter to tune symmetry. Feature-based methods do utilize more semantic knowledge than edge counting measures, but they are limited by their coverage and have a large dependency on tuning parameters according to the input ontology.

2.3.4 Information Content measures

To address the limitations of path based measures, Resnick (1995) [411] introduced the information content (IC) of concepts by incorporating corpus-based statistics into points of ontological graphs. The information content is the specificity of each concept and is calculated from the concept frequency in a large corpus of text. The IC of a term is the negative log of the probability of its occurrence:

$$IC_{corpus}(c) = -log(p(c)) = -log(\frac{freq(c)}{freq(root)}) \tag{2.13}$$

$$freq(c) = freq(c, C) + \sum_{c_s \epsilon children(c)} freq(c_s)$$

where $freq(c, C)$ is the number of times a concept c occurs in a corpus C, freq(c) is recursively defined as the frequency of concept c in addition to the frequencies of its children. By definition, concepts with higher frequencies are less informative and result in a lower IC. Resnick defines the notion of semantic similarity by the shared information between two terms. The Most Informative Common Ancestor (MICA) is the concept which subsumes the two concepts with maximum IC. The equation for semantic similarity proposed by Resnick is:

$$sim_{res}(c_1, c_2) = IC(MICA(c_1, c_2)) \tag{2.14}$$

The criticism for Resnick's measure is that many concepts that differ taxonomically may share the same MICA and will result with identical semantic similarity. Subsequent measurements incorporate the information content of the compared concepts. Lin (1998) [297] proposed the semantic similarity as a ratio between the shared IC between both terms and their individual IC:

$$sim_{lin}(c_1, c_2) = \frac{2 \cdot IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{2.15}$$

Jiang and Conrath (1997) [240] define the semantic distance (inverse of similarity) between two terms as the difference between the information content of each individual concept and the information content of their subsumer:

$$dist_{jcn}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \cdot IC(MICA(c_1, c_2)) \tag{2.16}$$

As Pedersen et al. [376] relates the similarity to the minimum path length with the sim_{path} measure, the Jiang and Conrath distance can be converted in the same manner $sim_{jcn} = \frac{1}{dist_{jcn}}$.

sim_{lin} and sim_{jcn} are relative measures between shared information and the information to fully describe each term. They do not take into account absolute measures and are displaced from the ontology graph. Schilicker et al. (2006) [445] proposes the relevance similarity measure which uses the probability of the MICA as a weighting factor:

$$sim_{Rel}(c_1, c_2) = sim_{lin} \times (1 - p(MICA(c_1, c_2)) \tag{2.17}$$

In most cases the MICA summarizes the information shared by both concepts, but concepts can have multiple inheritances and MICA may only capture the information partially. To accommodate for this limitation, Couto et al. [108, 109] proposed the GrASM and DiShIn strategies to aggregate the IC from their disjoint common ancestors. Disjoint common ancestors (DCA) are common ancestors that do not subsume any other common ancestors. It replaces the IC of the MICA with the average IC of all contributing DCA.

2.3.4.1 Intrinsic IC measures

Information content measures depend on annotated corpora which can be difficult to obtain, time consuming, and often not reliable. Intrinsic IC calculus models estimate the IC of concepts only by considering the structural information extracted from a given ontology. Seco et al. [450] define the intrinsic IC of a concept as a function of the number of its descendants.

$$IC_{seco}(c) = 1 - \frac{log(hypo(c) + 1)}{log(hypo(root))} \tag{2.18}$$

where hypo(c) is the number of hyponyms, or descendants, of concept c, hypo(root) is the number of descendants from the root node. However, this measure does not differentiate between different levels of generality. Concepts with equal amounts of descendants are treated equivalently despite how deeply located they are in taxonomy graphs. Zhou et al. [547] augment IC_{seco} by incorporating the relative depth of each concept in the ontology.:

$$IC_{zhou}(c) = k(IC_{seco}) + (1-k)(\frac{log(depth(c))}{log(max\_depth)} \tag{2.19}$$

where k adjust the weights of the two features. k = 0.5 was used in the study by Zhou et al. IC_{zhou} and IC_{seco} have been applied directly on the corpus-based IC measures (e.g. sim_{res}, sim_{lin} and sim_{jcn}) and have given comparable and superior results applied to Wordnet [434]. However, they will not properly differentiate IC without a largely well-defined ontology and detailed taxonomic structure. Concepts become too similar in a graph structure with minimal depth and a low branching factor. Sanchez et al [433] developed an IC measure that has less dependency on taxonomy design and considers multiple inheritances. It measures a ratio between the number of leaves and number of subsumers of a concept:

$$IC_{sanchez} = -log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \tag{2.20}$$

where leaves(c) is the number of leaves under concept c, subsumers(c) is the set of ancestors of c including itself, and max_leaves is the total amount of leaves in the taxonomy as a normalizing factor.

2.3.4.2 Redefining edge-counting measures in terms of IC

The advantages of edge-counting measures is there simplicity and low computational cost, but they omit substantial knowledge from the taxonomy by only focusing on path length. In an attempt to overcome these limitations, edge-counting measures can be redefined in terms of the IC of the concepts [432]. The minimum path is defined as the sum of differential information between two concepts, which simplifies to the Jiang and Conrath distance measure $dist_{jcn}$:

$$\begin{aligned} min_{\forall i}\,|path_i(c_1,c_2)| &\cong (IC(c_1)-IC(c_2))+(IC(c_2)-IC(c_1)) \\ &= (IC(c_1)-IC(lcs(c_1,c_2)))+(IC(c_2)-IC(lcs(c_1,c_2))) \\ &= IC(c_1)+IC(c_2)-2\times IC(lcs(c_1,c_2) \\ &= dist_{jcn}(c_1,c_2) \end{aligned} \tag{2.21}$$

The depth of concepts corresponds to the minimum path of the concept to the root node:

$$\begin{aligned} depth(c) &= min_{\forall i}\,|path_i(c,root)| \\ &\cong IC(c)+IC(root)-2\times IC(lcs(c,root)) \\ &= IC(c)+IC(root)-2\times IC(root) \\ &= IC(c)-IC(root) \\ &\cong IC(c) \end{aligned} \tag{2.22}$$

Generally $IC(root)$ can be viewed as 0, because it can subsume any concept, and it evaluates to zero when computed intrinsically. Therefore, the max_depth (used by sim_{lch}) may also be interpreted as the maximum IC (max_ic) held by any concept. To conclude this section, the following edge-counting measures are redefined with these equivalencies in terms of IC:

$$sim^{ic}_{path}(c_1,c_2) \cong \frac{1}{dist_{jcn}(c_1,c_2)} \tag{2.23}$$

$$sim^{ic}_{lch}(c_1,c_2) \cong -log\left(\frac{dist_{jcn}(c_1,c_2)+1}{2\times max\_ic}\right) \tag{2.24}$$

$$sim^{ic}_{wp} = \frac{2\times IC(MICA(c_1,c_2))}{IC(c_1)+IC(c_2)} \tag{2.25}$$

CHAPTER III

INTEGRATING BIOLOGICAL DATA

3.1 Introduction

Biologists have the unique challenge to classify a wealth of information that is constantly growing and evolving. Each new scientific discovery and laboratory advancement can yield new concepts and revise data representation. Resource providers have the potential to only focus on their own sub-discipline and philosophy of information structure. The annual Nucleic Acids Research journal database issue has listed over 1,552 database in 2014 covering molecular and cell biology [147]. Pertinent biological data is frequently stored across many different databases, hindering investigators that require information from multiple experiment types. It is unrealistic in todays climate to ask researchers to visit numerous databases with a point and click interface [182]. They are often confronted with this environment to answer common biological questions. To retrieve all necessary information requires an integrated system that can query multiple databases and minimize overhead. This chapter introduces a universal parser and relational schema translator that can be utilized for all NCBI databases in Abstract Syntax Notation (ASN.1) and all biological databases in Extensible Markup Language (XML).

These tools facilitate research scientists to locally integrate biological databases without significant workload or development time, negating many of the shortcomings experienced from data warehousing. All data is centralized into a local schema where it can be cleaned, filtered, or annotated according to the needs of the project. Queries are expedited without a reliance on network connectivity and are given fine granular access to the data [453]. Any revisions to the data models by the sources do not require continuous schema revisions and source code modifications.

Biological **S**earch **P**ropagate **I**ntegrate **D**emonstrate **A**nalyze, BioSPIDA, is a utility that helps alleviate the shortcomings of local integration. It contains a universal parser and relational translator written in Python and PostgreSQL that can be used for all biological databases in ASN.1 or XML format. The software does not need to be updated upon changes to the source data structure. Researchers are not required to develop a separate parser for each individual database they wish to integrate. BioSPIDA provides compelling features that significantly reduce storage requirements and implementation time to create a fully integrated system. Streaming of XML files into segments eliminate the necessity to initially load data into Document Object Models. This moderates memory requirements, especially in cases where file sizes can reach up to gigabytes of data. Parallelization of the parsing module allows biological databases to be loaded in fractions of time compared to its execution as an individual process. Storage requirements are minimized by a redundancy removal feature that identifies and removes all data duplications. These features significantly shrink the bottlenecks required to parse, transform, and load large biological databases into a locally integrated relational database.

3.2 Related Work

To meet the demands of the complexity and diversity of the biological community, NCBI has chosen to use Abstract Syntax Notation (ASN.1) [368]. It is a hierarchical data model that contains records, collections and variant types [208]. Its flexibility can handle needs for constant database restructuring where a relational system would be unsuccessful. ASN.1 is equally expressive as it is flexible. The semantics and biological meaning of the data can be interpreted without additional constraints of understanding a relational model. The standardized format allows exchange of information without dependence on a specific hardware or software environment [325].

As ASN.1 gives many advantages to the scientific community, it creates several limitations for procedures that require a broad range of investigation. Full text records need post-processing and parsing of their results. Scientists may spend 80% of their time retrieving, manipulating and preparing their data for analysis [414].

The biological sciences contain many diverse data types and it is difficult to create a global schema that precisely captures all necessary features. XML is a flexible and extensible markup language that is ideal in modelling structural heterogeneity and it is provided by many Web Services in Mediator based approaches [522], such as NCBI [442], European Molecular Biology Laboratory (EMBL) [518] and DNA Data Bank of Japan [168]. Despite the advantages of XML, it still has key limitations when compared to relational schemas. It is difficult to model complex properties that are required to represent biological pathways and many-to-many relationships [304].

There are many approaches to database integration each with their own advantages and disadvantages as described in Chapter 2. Data warehousing extracts and integrates information into a local relational database. SQL-level queries give fine granular access increasing the control, speed and global reach of analysis. Data warehouses are most appropriate when performance, local control, and privacy are essential and have facilitated studies that normally could not be performed under other conditions [304]. Mootha et al was able to identify one of the disease genes implicated in Leigh syndrome using integrated sequence and annotation data [341]. Shah et al helped determine yeast orthologs for human genes implicated in disease with sequence, taxonomy and genetic disorder databases [453]. Data warehouses can create a powerful system that delivers effective results, but they require extensive knowledge in database development as well as considerable setup-time to import each data source. BioSPIDA is a universal parser and relational schema translator for the biological databases with enhanced features of memory management, parallelization, and redundancy removal. In the following sections, each of its components are described to effectively illustrate how it can alleviate many of the shortcomings of data warehousing.

3.3 BioSPIDA

BioSPIDA is a utility that helps reduce the workload and implementation requirements for developing a fully integrated data warehouse. It contains a universal parser and relational translator written in Python and PostgreSQL that can be used for all NCBI databases in ASN.1 and biological databases in XML format. This allows researchers to create a local schema without detailed domain knowledge of each database selected for integration. The universal parser only requires a template file of all ASN.1 data types provided by NCBI or an XSD or DTD document of the biological provider's data model. It automatically determines how to parse each database without custom source code. The relational translator converts all data types into a fully representative relational schema that can be accessed with SQL-level queries for greater specificity of results. The details of these methods are described in the following sections.

3.3.1 Abstract Syntax Notation

ASN.1 objects can have many attributes that all have different data types. These include primitive, hierarchical, nested, enumerated, variant, and collection types. Primitive types hold basic values, such as integers, real numbers, booleans, and strings of characters. Enumerated types restrict attributes to a list of possible values. Variant types allow attributes to be many different data types. Attribute types can also be ASN.1 objects themselves creating a hierarchical structure. Keywords are used to identify the data type of each attribute, such as Integer, VisibleString, REAL, BOOLEAN, and ENUMERATED. CHOICE is used for variant data types.

Optional modifiers can be used for attributes that further change their data types.

The SEQUENCE OF or SET OF keyword treats the attribute as a collection. A sequence is an ordered collection, while a set does not enforce the order. Authors of a publication are an example of a sequenced collection, and MeSH terms are an example of a set collection data type. OPTIONAL and DEFAULT keywords flag attributes specifying if they are required or have a default value.

ASN.1 objects are rarely a simple tuple of attributes with primitive data types (Figure 3.1). The data type of an attribute can be an ASN.1 object linking itself to another tuple of attributes (Figure 3.2). Attributes can also be collections of values. It is possible for data types of a collection to be lists of primitive values or nested ASN.1 objects (Figure 3.3). Inside ASN.1 objects, variants may permit multiple data types for one attribute (Figure 3.4). Attributes also have the ability to be a tupled nest of child attributes (Figure 3.5). There is not a large set of restrictions for ASN.1 objects and they can be primitive, enumerated or variant data types themselves (Figure 3.6).

```
Textannot-id ::= SEQUENCE {
    name             VisibleString OPTIONAL ,
    accession  VisibleString OPTIONAL ,
    release    VisibleString OPTIONAL ,
    version    INTEGER       OPTIONAL}
```

Figure 3.1: ASN.1 object with only primitive attributes

```
Patent-seq-id ::= SEQUENCE {
    seqid INTEGER,
    cit Id-pat }

Id-pat ::= SEQUENCE {
    country VisibleString, ...,
    doc-type VisibleString OPTIONAL }
```

Figure 3.2: ASN.1 object with attribute linking to a child object type

```
Packed-seg ::= SEQUENCE {
    dim INTEGER DEFAULT 2,
    numseg INTEGER,
    ids SEQUENCE OF Seq-id,
    starts SEQUENCE OF INTEGER, ...}
```

Figure 3.3: ASN.1 object with collection attributes of primitive and object types

```
Score ::= SEQUENCE {
    id Object-id OPTIONAL ,
    value CHOICE {
        real REAL ,
        int INTEGER }
```

Figure 3.4: ASN.1 object with variant attribute permitting multiple data types

```
Int-fuzz ::= CHOICE {
     p-m INTEGER,
     range SEQUENCE {
          max INTEGER,
          min INTEGER },
     pct INTEGER ... }
```

Figure 3.5: ASN.1 object with attribute with nested tuple of child attributes

```
Na-strand ::= ENUMERATED {
     unknown (0),
     plus (1),
     minus (2), ...,
     other (255) }

PDB-mol-id ::= VisibleString

Product-pos ::= CHOICE {
     nucpos INTEGER,
     protpos Prot-pos }
```

Figure 3.6: ASN.1 objects that are primitive, enumerated and variant types

The flexibility in the ASN.1 data structure leads to elaborate scenarios yielding many distinct cases. The following are examples illustrating the possible complexity. An ASN.1 object can be a variant type that is itself a collection (Figure 3.7). An attribute can be a collection of ASN.1 objects, where the ASN.1 object derives itself into an enumerated data type (Figure 3.8). One of the data types of a variant attribute can be a tupled list of child attributes (Figure 3.5).

```
Title ::= SET OF CHOICE {
    name VisibleString, ...,
    issn VisibleString,
    abr VisibleString,
    isbn VisibleString }
```

Figure 3.7: ASN.1 object that is a collection of variant types

```
Dense-seg ::= SEQUENCE {
    dim INTEGER DEFAULT 2,
    numseg INTEGER,
     strands SEQUENCE OF Na-strand, ... }
```

Figure 3.8: ASN.1 object with collection attribute of object type that derives into enumerated type. Na-strand is an object that is an enumerated type (Figure 3.6).

3.3.2 Universal Parser

To build a universal parser for the NCBI databases, a context-free grammar must be constructed that successfully handles all data types provided by NCBI. Since each database uses the ASN.1 object model, the grammar must cover all possible types that are supported by the notation. This leads to an extra layer of complexity, because data types allow nested combinations that develop into numerous unique cases. The context-free grammar for ASN.1 notation used by BioSPIDA can be seen in Figure 3.9. The advantage of using a context-free grammar is that NCBI databases can be selected for download without having to create an individual parser. The universal parser in BioSPIDA uses the grammar to read ASN.1 specification files, automatically learning how to read each database.

$\langle \mathit{ASN} \rangle$ ::= word ‘`::=`’ $\langle \mathit{attributeType} \rangle$ | $\langle \mathit{ASN} \rangle$ $\langle \mathit{ASN} \rangle$

$\langle \mathit{attributeType} \rangle$ ::= $\langle \mathit{multi} \rangle$ $\langle \mathit{variableType} \rangle$ $\langle \mathit{default} \rangle$ $\langle \mathit{optional} \rangle$

$\langle \mathit{multi} \rangle$ ::= ‘`set of`’ | ‘`sequence of`’

$\langle \mathit{variableType} \rangle$ ::= word | $\langle \mathit{enumeratedType} \rangle$ | $\langle \mathit{choiceType} \rangle$

$\langle \mathit{enumeratedType} \rangle$ ::= ‘`enumerated {`’ $\langle \mathit{enumeratedList} \rangle$ ‘`}`’

$\langle \mathit{enumeratedList} \rangle$ ::= word ‘`(`’ number)’ | $\langle \mathit{enumeratedList} \rangle$ ‘`,`’ $\langle \mathit{enumeratedList} \rangle$

$\langle \mathit{choiceType} \rangle$ ::= ‘`choice`’ $\langle \mathit{choiceTypeRight} \rangle$ | ‘`sequence`’ $\langle \mathit{choiceTypeRight} \rangle$

$\langle \mathit{choiceTypeRight} \rangle$ ::= ‘`{`’ $\langle \mathit{attributeList} \rangle$ ‘`}`’

$\langle \mathit{attributeList} \rangle$ ::= $\langle \mathit{attribute} \rangle$ | $\langle \mathit{attributeList} \rangle$ ‘`,`’ $\langle \mathit{attributeList} \rangle$

$\langle \mathit{attribute} \rangle$ ::= word $\langle \mathit{attributeType} \rangle$

$\langle \mathit{default} \rangle$ ::= ‘`default`’ $\langle \mathit{value} \rangle$ | ϵ

$\langle \mathit{optional} \rangle$ ::= ‘`optional`’ | ϵ

$\langle \mathit{value} \rangle$ ::= word | number

Figure 3.9: Context-free grammar for ASN.1

BioSPIDA generates ASN.1 template files for each targeted NCBI database by reading the NCBI ASN.1 specifications file available online[1] that contains all supported data objects. The parent ASN.1 object for the targeted NCBI database must be supplied to BioSPIDA to create the ASN.1 template file. For example, the parent objects for the following databases are EntrezGene (Entrez Gene), Pubmed-entry

[1] http://www.ncbi.nlm.nih.gov/data_specs/asn/NCBI_all.asn

(Pubmed), Bioseq-set (GenBank), Bioseq-set (RefSeq), Biostruc (MMDB), CDD (CDD). If the data structure changes for a given NCBI database, it is only necessary to redownload the updated NCBI ASN.1 specifications file. There is no obligation for software code changes. When initially selecting a database for conversion, researchers must specify whether the uncompressed ASN.1 xml data is locally stored or requires remote retrieval. If remote retrieval is necessary, BioSPIDA requires a text file of database records ids to be sent to the NCBI EFetch utility [440].

NCBI provides access to each of their databases either by their ftp site[2] or their EFetch utility. The data for each record is available in xml format and it is structured according to the document type definition file[3]. The XML tags identify each possible attribute and item given for each record. The software code in BioSPIDA does not require identifying each individual xml tag to search for when parsing database records. All that is required is the ASN.1 specification file describing the data model of the target database. BioSPIDA has the ability to learn how to parse the xml from each NCBI database without human intervention, because NCBI follows a strict set of guidelines when converting their ASN.1 records to xml format. NCBI presents XML for each database record that follows the exact structure given by their ASN.1 data model[4].

Figures 3.10 and 3.11 compare the ASN data model for EntrezGene with example xml retrieved from a database record. The parent xml tag is <Entrezgene> which is the parent ASN.1 object in the definition file. An xml tag is created for each attribute of the ASN.1 object. <Entrezgene_type> conveys that the "Entrezgene" object has an attribute named "type." Since the attribute is an INTEGER, it is a primitive type. Therefore, its value is placed immediately after its identifier.

The data type for the track-info attribute is another ASN.1 object, Gene-track.

[2] https://www.ncbi.nlm.nih.gov/Ftp/

[3] http://www.ncbi.nlm.nih.gov/dtd/

[4] http://www.ncbi.nlm.nih.gov/data_specs/asn/

The value for the track-info attribute cannot be placed immediately after its xml tag. It must identify it is an ASN.1 object. The <Gene-track> xml tag follows the <Entrezgene_track-info> tag. The same logic is used to recursively list all of the subsequent attributes and child objects.

```
Entrezgene ::= SEQUENCE {
  track-info Gene-track,
  type INTEGER,
  source BioSource, ... }

Gene-track ::= SEQUENCE {
   geneid VisibleString,
   create-date Date, ... }

BioSource ::= SEQUENCE {
    org Org-ref ... }

OrgName ::= SEQUENCE {
   name CHOICE {
      binomial BinomialOrgName
      virus VisibleString
      hybrid MultiOrgName ... }
```

Figure 3.10: ASN.1 specification for Entrezgene

```
<Entrezgene>
    <Entrezgene_track-info>
         <Gene-track>
            <Gene-track_geneid>1246500
      <Gene-track_create-date>
          <Date>
          ...
    <Entrezgene_source>
        <BioSource>
           <Biosource_org>
               <Org-ref>
             <Org-ref_orgname>
               <OrgName>
             <OrgName_name>
               <OrgName_name_binomial>
                     ...
     <Entrezgene_type>6
</Entrezgene>
```

Figure 3.11: XML for Entrezgene record

If an attribute type is CHOICE, it can have many possible data types. An additional xml tag distinguishes the variant attribute. Below the <Orgname> xml tag, <Orgname_name> and <Orgname_name_binomial> are listed to identify the attribute and the chosen data type.

NCBI follows a consistent list of guidelines for generating xml of collection attributes in ASN.1 object records. Collections of primitive and variant attributes add an “_E” in the xml tag name to identify each entry. Variant types additionally add the chosen attribute name after the “_E” suffix. Collections of object types do not add the “_E” suffix in their xml tags. Instead, they have an open and closing tag of the object name for each item in the list. Figure 3.12 gives an example for these cases. GB-block contains the keywords attribute that is a collection of VisibleStrings. Seq-align has a collection of Score objects. The names attribute in the Auth-list object is a variant with one of its permitted types being a collection of VisibleStrings. Title is a collection of variant types. The generated xml for these objects all adhere to the same conventions (Figure 3.13).

```
GB-block ::= SEQUENCE {
   source VisibleString OPTIONAL,
   keywords SEQUENCE OF VisibleString OPTIONAL, ... }

Seq-align ::= SEQUENCE {
    score SET OF Score OPTIONAL, ... }

Score ::= SEQUENCE {
    id Object-id OPTIONAL ,
    value CHOICE {
        real REAL ,
        int INTEGER } }

Title ::= SET OF CHOICE {
    name VisibleString ,
    iso-jta VisibleString ,
    ml-jta VisibleString ,
    issn VisibleString , ... }

Auth-list ::= SEQUENCE {
   names CHOICE {
      std SEQUENCE OF Author,
      ml SEQUENCE OF VisibleString ,
      str SEQUENCE OF VisibleString },
   affil Affil OPTIONAL }
```

Figure 3.12: ASN.1 specification for collection types

```
<GB-block>
   <GB-block_source>Bradyrhizobium japonicum
   <GB-block_keywords>
    <GB-block_keywords_E>NifA dependance
    <GB-block_keywords_E>sigma factor 54
    <GB-block_keywords_E>transcriptional activator
   </GB-block_keywords>
</GB-block>

<Seq-align>
 <Seq-align_score>
  <Score>
  <Score_id>
    <Object-id>
    <Object-id_str>score
    </Object-id>
  </Score_id>
  <Score_value>
    <Score_value_int>2813
  </Score_value>
  </Score>
  <Score>
  <Score_id>
    <Object-id>
    <Object-id_str>bit_score
</Object-id_str>
    </Object-id>
  </Score_id>
  <Score_value>
    <Score_value_real>2573.13
  </Score_value>
  </Score>
 </Seq-align_score>
</Seq-align>

<Auth-list>
  <Auth-list_names>
  <Auth-list_names_ml>
    <Auth-list_names_ml_E>Brat DJ
    <Auth-list_names_ml_E>Castellano-Sanchez A
    <Auth-list_names_ml_E>Kaur B
  </Auth-list_names_ml>
  </Auth-list_names>
</Auth-list>

<Title>
  <Title_E>
  <Title_E_issn>0042-6822
  </Title_E>
  <Title_E>
  <Title_E_name>Virology
  </Title_E>
</Title>
```

Figure 3.13: XML for Entrezgene Record with collection types

The universal parser utilizes the exact rules that NCBI implements to deliver xml data of their database records. No software revisions are necessary for each additional NCBI database that must be loaded and parsed. Scientists are only required to specify the database name, and BioSPIDA creates a new template file that contains all of the objects used by the selected database.

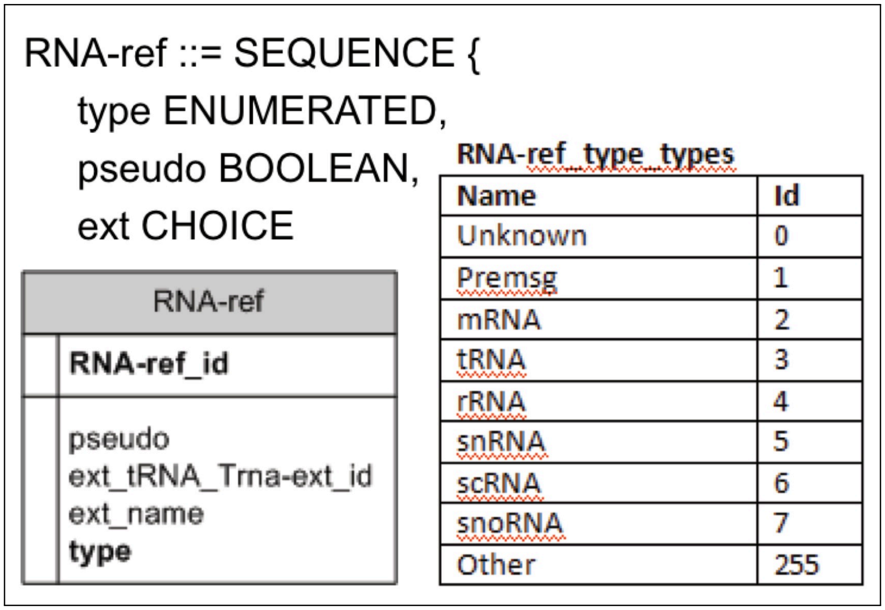
3.3.3 Relational Translator

Parsers only facilitate retrieval of the data. The information must still be inserted, stored, and analyzed. BioSPIDA contains a relational translator that automatically converts the ASN.1 data model from the given template file into a MySQL or Postgres relational database.

ASN.1 has a broad range of data types, and the translator must also follow a set of rules to create all tables and subsequent fields. Tables are created for every ASN.1 object specified in the template file. Attributes that are primitive data types are directly converted into fields. Enumerated types are also converted into fields as an integer data type. Separate reference tables are generated to list all of the values for each enumerated type. For variant types, individual fields are created for each possible data type. All of the variant attributes fields will be null for the cases not chosen.

A primary column is added to every table to uniquely identify each row inserted into the database. The value of the primary key is incremented for each record insertion. These identifying fields also exist in child tables as foreign keys facilitating table linkages. Separate linking tables are produced for collection types. The linking table includes a sequence field for SEQUENCE collections, but not for SET collections. Example tables generated from the relational translator for the Entrezgene database can be viewed in Figures 3.14-3.16.

Figure 3.14 displays the tables created for the RNA-ref object and the "type" attribute. Since the attribute is enumerated, a separate table named RNA-ref_-type_types is produced with all possible values.



```
RNA-ref ::= SEQUENCE {
    type ENUMERATED,
    pseudo BOOLEAN,
    ext CHOICE
```

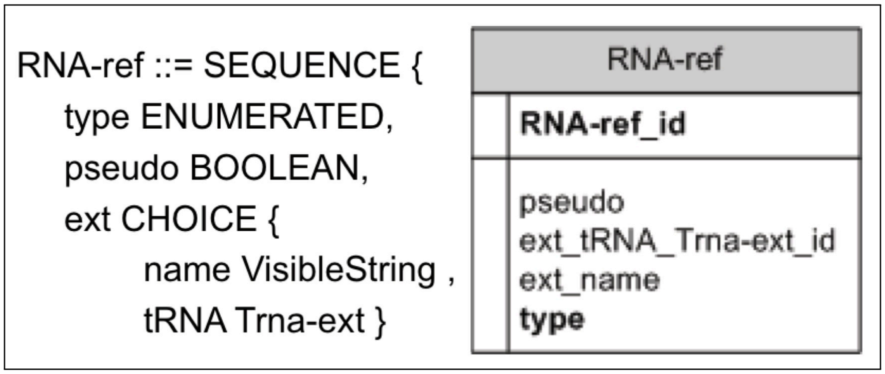
RNA-ref
RNA-ref_id
pseudo ext_tRNA_Trna-ext_id ext_name **type**

RNA-ref_type_types

Name	Id
Unknown	0
Premsg	1
mRNA	2
tRNA	3
rRNA	4
snRNA	5
scRNA	6
snoRNA	7
Other	255

Figure 3.14: Relational conversion of ASN.1 enumerated types

Figure 3.15 shows an example of a variant type. The ext attribute of the RNA-ref object can have multiple data types. If ext is a VisibleString, it is identified as a name. If it is a Trna-ext object, it is identified as tRNA. Therefore, there are fields in the RNA-ref table for ext_tRNA and ext_name. Only the field for the selected data type will not be null.

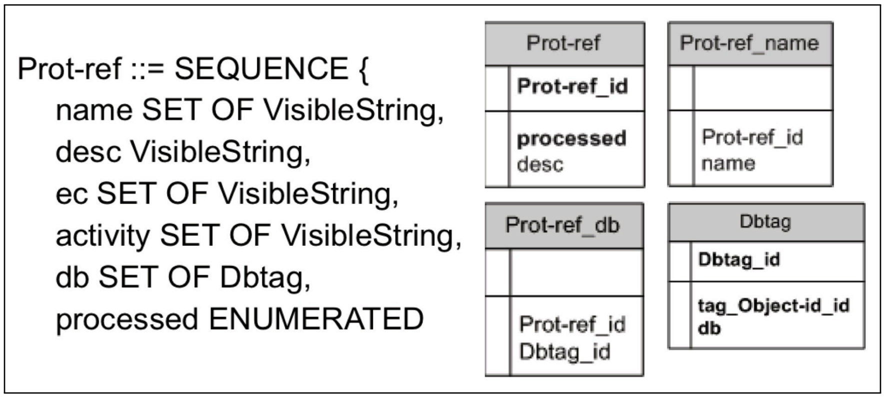


```
RNA-ref ::= SEQUENCE {
    type ENUMERATED,
    pseudo BOOLEAN,
    ext CHOICE {
          name VisibleString ,
          tRNA Trna-ext }
```

RNA-ref
RNA-ref_id
pseudo ext_tRNA_Trna-ext_id ext_name **type**

Figure 3.15: Relational conversion of ASN.1 variant types

In Figure 3.16, the Prot-ref object has attributes that are collection types. A separate linking table is generated for the "name" attribute labeled Prot-ref_name. Since, the attribute is a VisibleString, only its value and Prot-ref identifier are needed in the linking table. The db attribute in the Prot-ref object is a collection of Dbtag objects. Therefore, the Prot-ref_db linking table contains identifier fields for the Prot-ref table and the Dbtag table. The tables are linked with the Prot-ref_id and Dbtag_id columns.



```
Prot-ref ::= SEQUENCE {
    name SET OF VisibleString,
    desc VisibleString,
    ec SET OF VisibleString,
    activity SET OF VisibleString,
    db SET OF Dbtag,
    processed ENUMERATED
```

Figure 3.16: Relational conversion of ASN.1 collection types

The universal parser inserts all of the data it retrieves into loading files. A separate loading file is designated for each table created from the Relational Translator. The Relational translator also produces the load and create table statements for MySQL or Postgres. After the loading files are fully populated, the only human intervention required is to execute the create and load statements to generate a fully converted relational database.

3.3.4 Efficient Memory Management

Initially BioSPIDA was only tested against the Pubmed, Entrez Gene and OMIM NCBI databases [202]. Each record in these databases contains a considerable amount of fields, but the total size usually does not exceed over one megabyte. Parsing these records did not require an efficient memory management system, because the entire

Document Object Model (DOM) for each xml file could be loaded seamlessly. This was not the case for Genbank, Refseq, and MMDB where records grow substantially in size. The xml file for the largest record in Genbank was over 5.5 gigabytes. This did not allow xml records to be fully loaded into memory.

xml.dom.minidom is a library included in Python 2.0 which is a lightweight implementation of the DOM interface[5] which would result in memory errors when applied against larger NCBI databases. Massive records restrict full reads of xml data and prevent memory storage of transformed ASN.1 objects. lxml is a powerful xml parsing library that facilitated parsing of more extensive xml files[6]. It supports reading data from xml files one line at a time, rather than initially loading all DOM data. BioSPIDA is implemented with the lxml library and the memory requirements for storing ASN.1 objects are greatly reduced.

For each line of xml, BioSPIDA adds the new data parsed to the current ASN.1 object stored in memory. ASN.1 objects have a nested data structure, and there can be many child attributes that are also ASN.1 objects [208]. After all xml data is received for a child ASN.1 object, it is no longer needed for memory storage. Each completed child ASN.1 object is removed from memory and written to associated tab-delimited loading files. This significantly reduces the memory requirements even for cases of very large xml file sizes.

Figure 3.17 compares the memory states of loading XML files in entirety with the xml.dom.minidom library versus streaming large XML files with the lxml package. Figure 3.17a is a shortened sample of an xml record from EntrezGene. Figure 3.17b shows the ASN.1 schema for the EntrezGene object with all referring child objects. Figure 3.17c depicts the memory state for the complete ASN.1 object after fully parsing the EntrezGene record. Figures 3.17a - 3.17c illustrate the transformation in

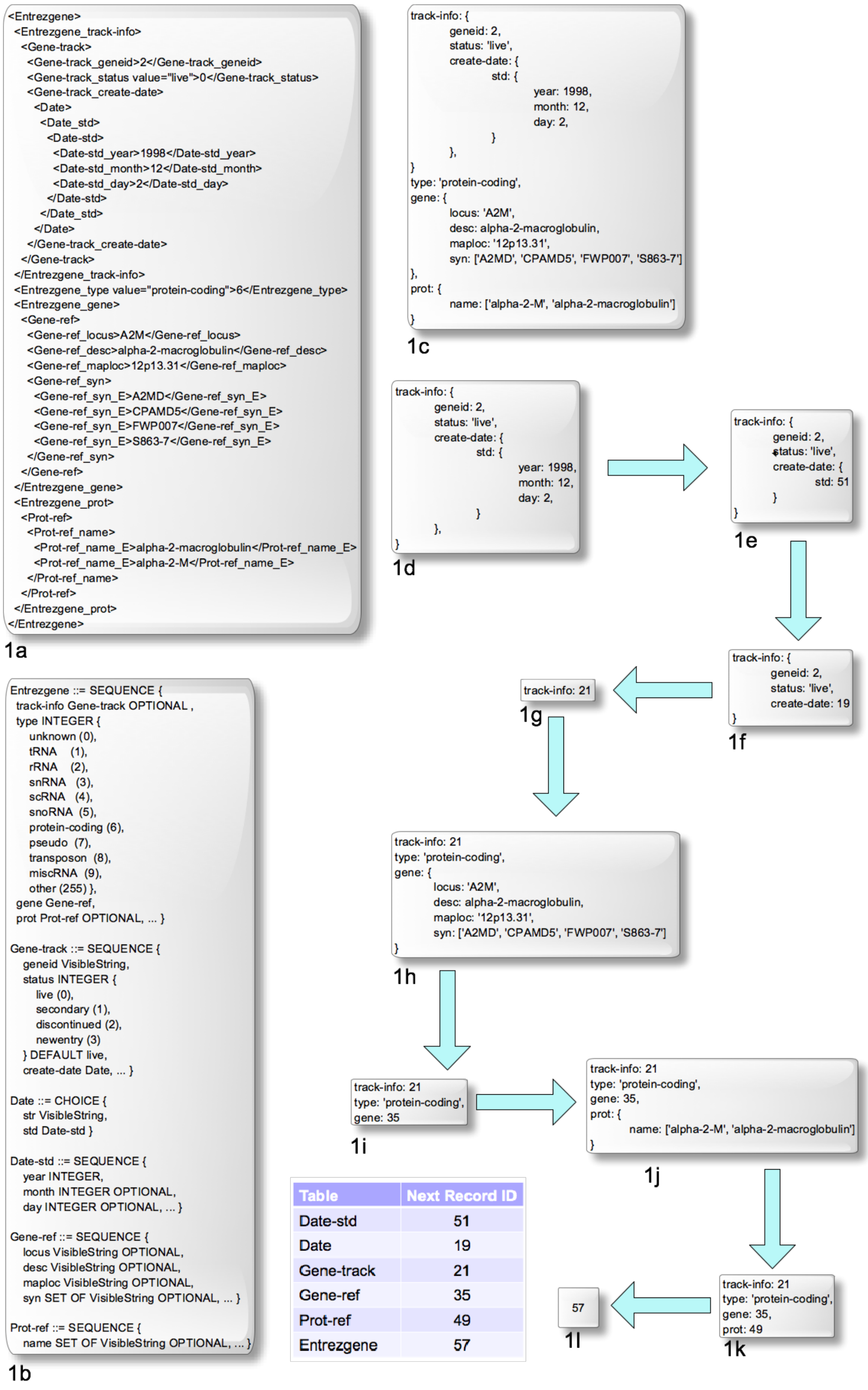
[5] http://docs.python.org/library/xml.dom.minidom.html#module-xml.dom.minidom/

[6] http://lxml.de

data structure as xml files are fully parsed and loaded into memory. The average size of EntrezGene records is only 24.8 Kb facilitating reading entire xml records without efficient memory management.

BioSPIDA 1.1 continually refreshes the memory stored by ASN.1 objects to support parsing of larger database records. Figures 3.17d - 3.17l represent the memory states of BioSPIDA while reading each line of XML. Initially the track-info attribute is loaded into the Entrezgene ASN.1 object. The data type of the track-info attribute is an ASN.1 object, Gene-track (Figure 3.17d). Gene-track has three attributes: geneid, status, and create-date. The data types for geneid and status are primitive types, VisibleString and Integer. The attribute create-date is the ASN.1 object Date. Date has only one attribute, std, which is the ASN.1 object Date-std. Date-std has 3 attributes, year, month and day.

After BioSPIDA has reached the closing xml tag for Date-std, it has completed parsing the Date-std ASN.1 object (Figure 3.17d). It is no longer needed to store the Date-std object in memory. The data is inserted into the loading file for the Date-std database table. After insertion, a unique record identifier from the Date-std table is returned. In this example, the value 51 is returned and the std attribute is updated in memory to only contain the identifier number (Figure 3.17e). The next lines in the xml file complete parsing of the Date and Gene-track objects. Figures 3.17f and 3.17g show the updated memory states. The rest of the xml file contains the type, gene, and prot attributes in the Entrezgene ASN.1 object. The prot and gene attributes are the ASN.1 objects Gene-ref and Prot-ref. Figures 3.17h - 3.17k are the memory states after the gene and prot attributes are updated after table insertion. After parsing has completed for the xml file, the parent Entrezgene ASN.1 object is inserted into the Entrezgene database table and a unique record identifier is returned (Figure 3.17l).



1a

```
<Entrezgene>
 <Entrezgene_track-info>
  <Gene-track>
   <Gene-track_geneid>2</Gene-track_geneid>
   <Gene-track_status value="live">0</Gene-track_status>
   <Gene-track_create-date>
    <Date>
     <Date_std>
      <Date-std>
       <Date-std_year>1998</Date-std_year>
       <Date-std_month>12</Date-std_month>
       <Date-std_day>2</Date-std_day>
      </Date-std>
     </Date_std>
    </Date>
   </Gene-track_create-date>
  </Gene-track>
 </Entrezgene_track-info>
 <Entrezgene_type value="protein-coding">6</Entrezgene_type>
 <Entrezgene_gene>
  <Gene-ref>
   <Gene-ref_locus>A2M</Gene-ref_locus>
   <Gene-ref_desc>alpha-2-macroglobulin</Gene-ref_desc>
   <Gene-ref_maploc>12p13.31</Gene-ref_maploc>
   <Gene-ref_syn>
    <Gene-ref_syn_E>A2MD</Gene-ref_syn_E>
    <Gene-ref_syn_E>CPAMD5</Gene-ref_syn_E>
    <Gene-ref_syn_E>FWP007</Gene-ref_syn_E>
    <Gene-ref_syn_E>S863-7</Gene-ref_syn_E>
   </Gene-ref_syn>
  </Gene-ref>
 </Entrezgene_gene>
 <Entrezgene_prot>
  <Prot-ref>
   <Prot-ref_name>
    <Prot-ref_name_E>alpha-2-macroglobulin</Prot-ref_name_E>
    <Prot-ref_name_E>alpha-2-M</Prot-ref_name_E>
   </Prot-ref_name>
  </Prot-ref>
 </Entrezgene_prot>
</Entrezgene>
```

1b

```
Entrezgene ::= SEQUENCE {
 track-info Gene-track OPTIONAL ,
 type INTEGER {
     unknown (0),
     tRNA    (1),
     rRNA    (2),
     snRNA   (3),
     scRNA   (4),
     snoRNA  (5),
     protein-coding (6),
     pseudo  (7),
     transposon  (8),
     miscRNA  (9),
     other (255) },
 gene Gene-ref,
 prot Prot-ref OPTIONAL, ... }

Gene-track ::= SEQUENCE {
   geneid VisibleString,
   status INTEGER {
      live (0),
      secondary (1),
      discontinued (2),
      newentry (3)
   } DEFAULT live,
   create-date Date, ... }

Date ::= CHOICE {
   str VisibleString,
   std Date-std }

Date-std ::= SEQUENCE {
   year INTEGER,
   month INTEGER OPTIONAL,
   day INTEGER OPTIONAL, ... }

Gene-ref ::= SEQUENCE {
   locus VisibleString OPTIONAL,
   desc VisibleString OPTIONAL,
   maploc VisibleString OPTIONAL,
   syn SET OF VisibleString OPTIONAL, ... }

Prot-ref ::= SEQUENCE {
   name SET OF VisibleString OPTIONAL, ... }
```

1c

```
track-info: {
        geneid: 2,
        status: 'live',
        create-date: {
                std: {
                        year: 1998,
                        month: 12,
                        day: 2,
                }
        },
}
type: 'protein-coding',
gene: {
        locus: 'A2M',
        desc: alpha-2-macroglobulin,
        maploc: '12p13.31',
        syn: ['A2MD', 'CPAMD5', 'FWP007', 'S863-7']
},
prot: {
        name: ['alpha-2-M', 'alpha-2-macroglobulin']
}
```

1d

```
track-info: {
        geneid: 2,
        status: 'live',
        create-date: {
                std: {
                        year: 1998,
                        month: 12,
                        day: 2,
                }
        },
}
```

1e

```
track-info: {
        geneid: 2,
        status: 'live',
        create-date: {
                std: 51
        }
}
```

1f

```
track-info: {
        geneid: 2,
        status: 'live',
        create-date: 19
}
```

1g

```
track-info: 21
```

1h

```
track-info: 21
type: 'protein-coding',
gene: {
        locus: 'A2M',
        desc: alpha-2-macroglobulin,
        maploc: '12p13.31',
        syn: ['A2MD', 'CPAMD5', 'FWP007', 'S863-7']
}
```

1i

```
track-info: 21
type: 'protein-coding',
gene: 35
```

1j

```
track-info: 21
type: 'protein-coding',
gene: 35,
prot: {
        name: ['alpha-2-M', 'alpha-2-macroglobulin']
}
```

1k

```
track-info: 21
type: 'protein-coding',
gene: 35,
prot: 49
```

1l

```
57
```

Table	Next Record ID
Date-std	51
Date	19
Gene-track	21
Gene-ref	35
Prot-ref	49
Entrezgene	57

Figure 3.17: XML Parsing a.) Entrez Gene xml record b.) ASN.1 schema for Entrez Gene object c.) Memory state parsing full xml record d.- l.) Memory states incrementally parsing a xml record

BioSPIDA continually reduces allocated memory after fully loading ASN.1 objects from xml files. This minimizes the memory requirements to negligible values, but also brings advantages for cases of abnormal program termination. It can take hours to parse xml files with massive file sizes reaching up to several gigabytes. Since BioSPIDA regularly inserts to loading files while parsing xml records, it can easily resume at the point of termination. Resume technology is imperative for Genbank and Refseq where xml files can reach enormous sizes.

3.3.5 Parallelization

NCBI databases contain details of nucleotide sequences, medical literature, gene expression, molecular structure and three-dimensional structures. An entire NCBI database can range from hundreds of gigabytes to several terabytes. Due to these high volumes of data, researchers typically only download subsets of databases for local integration. BioSPIDA attempts to reduce setup overhead by supporting the download and parsing of NCBI databases on parallel processors. This does not slow each individual process, because they do not have to be tightly synchronized. Each process communicates with a Shared Server that contains the next record id number for each table. The Shared Server also maintains a queue for each xml record to be parsed. This prevents individual processes from parsing identical xml records and from using redundant identifier keys during table insertion.

The parallel programming model utilized by BioSPIDA is a Single Instruction Multiple Data (SIMD) architecture [155]. All tasks execute the same program simultaneously, but operate on different data. Parallel processes are executed within a local area network (LAN) with machines in close physical proximity. BioSPIDA uses a message-passing model where each task uses its own local memory. Running processes only send and receive messages to a shared server. The shared server is a dedicated program to handle read and write requests and communicates asynchronously. It is unaware when updates are needed and each running process must explicitly send requests to receive data. The shared server contains global information of the current table row ids and the remaining xml records to be parsed. During communication, a locking mechanism is used to ensure table ids are updated before sending data.

Even though BioSPIDA utilizes a universal parser for each database, minimal implementation was required to support parallel processing. The ASN.1 records for each database are stored differently at the NCBI ftp site[7]. EntrezGene contains one large file for all records, while Genbank, Refseq, MMDB, CDD contain a separate file for each record. Gene2xml and datatool are utilities provided by the NCBI Toolkit [506] that are used to convert each ASN.1 record into xml. The Shared Server contains a queue that holds the file names for each ASN.1 data record (Figure 3.18). For EntrezGene, the large file is initially split into parts for each parallel processor.

The data files are not available by ftp for every NCBI ASN.1 database. Pubmed currently has over 24 million records, and NCBI does not provide the database for download. However, the EFetch utility allows retrieval of specific database records when given an associated identifier number[8]. EFetch supports PubMed as well Entrez Gene, CDD, MMDB, BioSystems, and others[9]. This is useful for cases where downloadable ftp data is unavailable or if it is preferred to only receive a subset of data. For the parallel implementation of Pubmed, the Shared Server contains a queue of id numbers rather than file names. These id numbers are then sent to EFetch and the resulting xml records are downloaded from NCBI (Figure 3.19).

[7] https://www.ncbi.nlm.nih.gov/Ftp/

[8] http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi

[9] http://www.ncbi.nlm.nih.gov/books/NBK25497/table/chapter2.chapter2_-table1/?report=objectonly

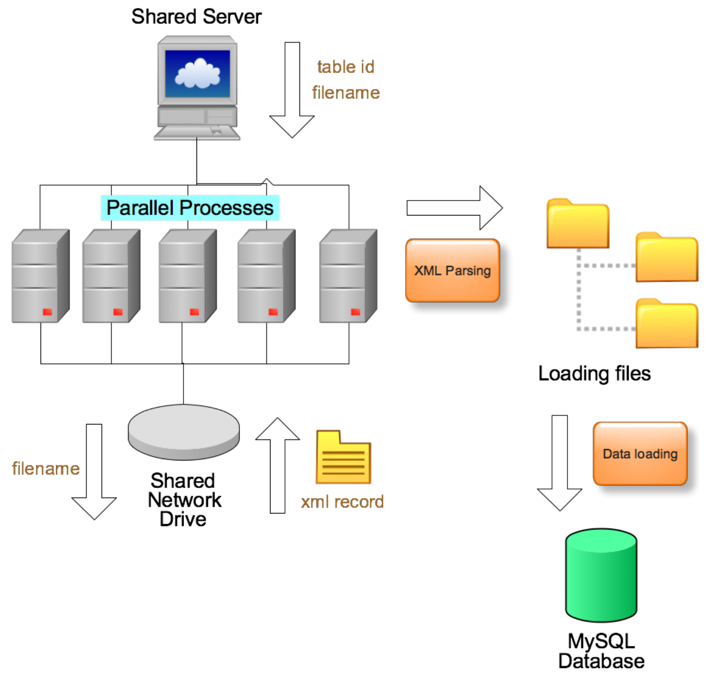


Figure 3.18: Parallel implementation of XML parsing process with ASN.1 records housed on Local Area Network.

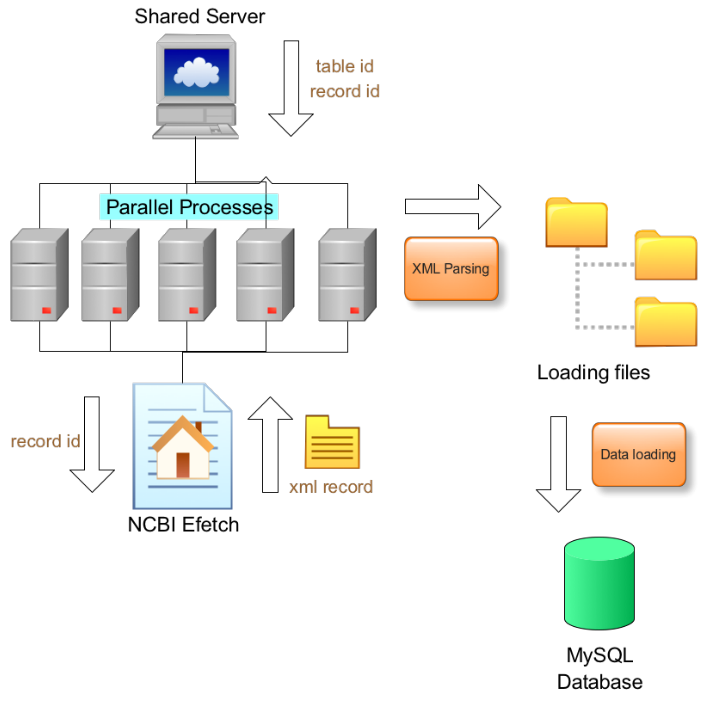


Figure 3.19: Parallel implementation of XML parsing process with ASN.1 records housed on external NCBI server.

Before adding support for parallel processing, a small header file must be created for each database. It specifies whether the xml records are housed locally or are retrieved online with EFetch. The header file provides a list of identifier numbers for remote records or a list of file names for local records. The Shared Server reads the list into its queue. Each parallel process receives a new id number or file name after completing its current task. The parallel process will download and parse a new xml record from EFetch if given an id number. If given a file name, the parallel process matches and parses the local xml file. All parallel processes are connected to a shared LAN, so there is no need to transfer files between machines if the xml records are housed locally.

3.3.6 Redundancy Removal

The new features of BioSPIDA are aimed at reducing the system overhead to implement a locally integrated data warehouse of NCBI ASN.1 databases. Streaming of large XML files and parallelization lower the memory requirements and setup time. However, larger NCBI databases can reach up to several terabytes in file sizes and can necessitate massive amounts of storage space. BioSPIDA attempts to reduce these requirements by removing any redundancy in the parsed data before insertion into the MySQL or Postgres database.

Each xml record in NCBI databases contain information from all relevant fields. A publication record in PubMed includes the title, abstract, all associated authors, and many other data items. All of this information is appended to tab-delimited loading files for each associated database table. This can create a lot of redundancy. It may be rare for a publication to have the same title or abstract, but it is much less rare to have identical authors or submission dates. The parsing process cannot check if an author or other entity item was used previously before inserting into its associated loading file. This would significantly slow execution time to search billions of rows before every insertion. It is very common to find duplicate records in loading files that each have unique row identifier numbers.

It can appear superficially that redundancy can easily be resolved by removing all duplicate records from each loading file. However, each loading file represents a table to be inserted into the database. Tables can have foreign keys that refer to other child tables. If duplicate records were removed from a child table, the parent table must update its foreign key values to refer to the same records. Redundancies from loading files cannot be removed in any order. They must occur in a strict order from child to parent tables.

It is difficult to manually determine the correct order for redundancy removal, because each converted relational database can contain hundreds of tables with many possible parent and child connections. A topological sorting algorithm [105] is used that automatically calculates the order to update each table. It returns a list where no table will be updated unless all of its child tables have been processed previously.

In Figure 3.20, the order is illustrated for removing redundancy for tables in Entrez Gene. Object-id has no children and it is the first table to remove all duplicate records. The tag_Object-id_id foreign key in Dbtag is updated with the new identifier values from Object-id after all redundant records are removed. Duplicate records can then also be removed from Dbtag, because it does not have any other child tables. The foreign keys in the tables Org-ref_db, Gene-ref_db, and Prot-ref_db are then updated with the new Dbtag identifier values. Redundancies cannot yet be removed from these tables, because they have many other child tables (not depicted in Figure 3.20). The next tables to be checked for duplicates are BinomialOrgName, OrgName, Org-ref, and Biosource. This process continues until redundancies are removed from every table and all foreign keys are updated in every parent table.

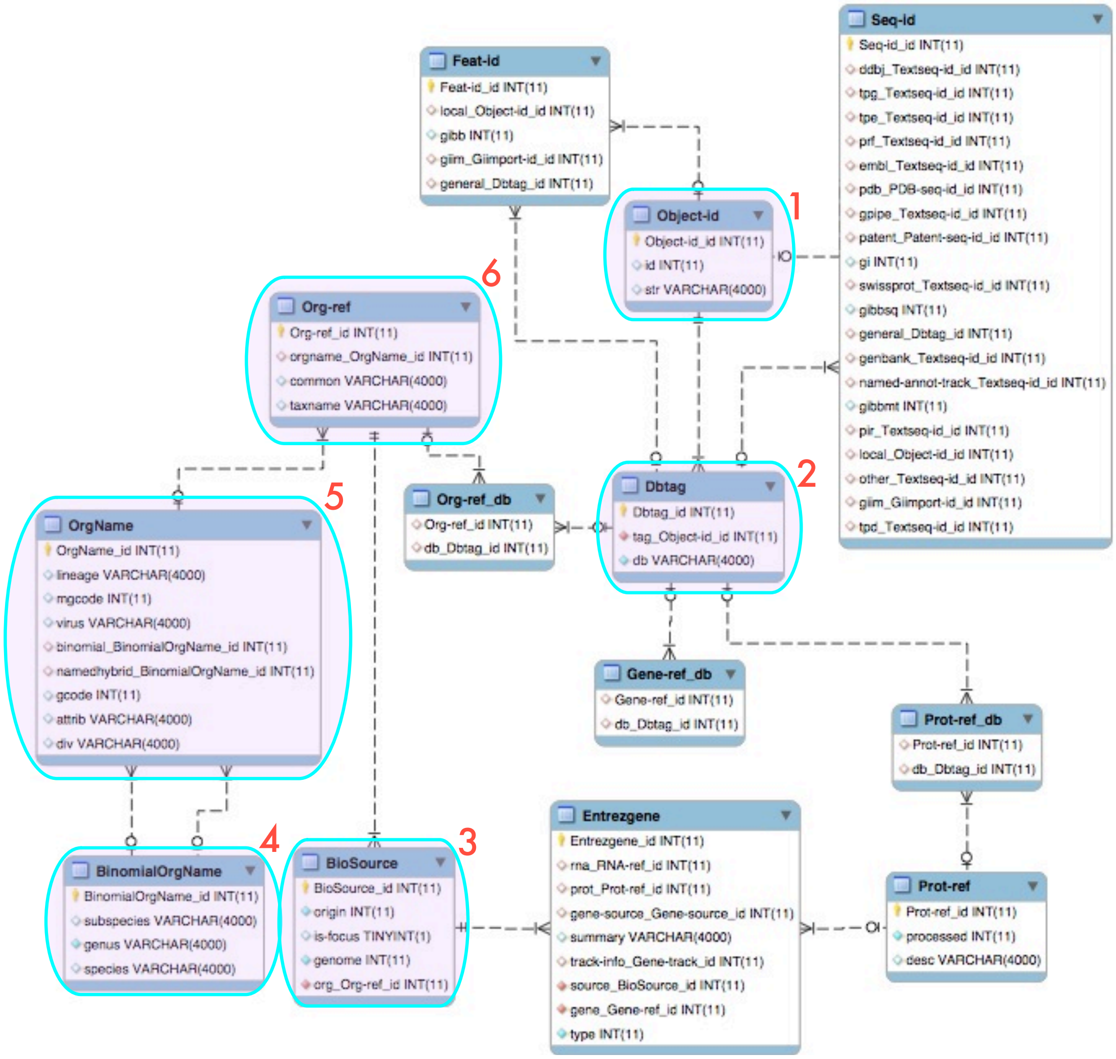


Figure 3.20: Order of tables that undergo redundancy removal according to list returned by topological sort.

There is an exception that BioSPIDA must take into account in determining the parents of each table in the relational schema. A parent of a table can be either an object table or a list table. An object table contains a primary key that uniquely identifies each row. List tables do not have primary keys, because they are only used for linkages. For example, Pubmed-entry contains a list of Pubmed-urls. Pubmed-entry_urls is a list table that connects the Pubmed-entry and Pubmed-url table. Therefore, Pubmed-entry_urls is a parent of Pubmed-entry and Pubmed-url, because it contains the identifier keys for both tables. If a natural topological sort were executed on this schema, Pubmed-entry_url would only be identified as dependent on these two child tables. However, Pubmed-entry cannot be sorted and undergo redundancy removal before Pubmed-url. Pubmed-entry contains a list of Pubmed-urls, and this cannot not be determined by only inspecting the Pubmed-entry table. The parents of Pubmed-url must be marked as Pubmed-entry_url as well as Pubmed-entry. This relationship can be observed in Figure 3.21a.

If a parent of an object table is a list table, the object table must be merged before it is sorted and redundancies are removed. In the case of Pubmed-entry, Pubmed-url would first identify its duplicate rows and update its parent table Pubmed-entry_url with its new identifier key values. Pubmed-entry would then merge with Pubmed-entry_url to create a temporary table with all its associated urls. The temporary table is then sorted to find and remove all its redundancies. A new Pubmed-entry table is then split from the temporary merge table with only unique entries. Pubmed-entry_url is again updated but now with the new identifier keys from Pubmed-entry. If there are any duplicate rows in Pubmed-entry_url after updating, they are removed as well. A simpler case is when an object table's parent is a list table, but it does not refer to any other tables. Medline-entry contains a list of pub-types. Its parent list table is Medline-entry_pub-type. Medline-entry_pub-type only contains the pub-type field which is a textual value that does not link to any other table. Medline-entry would still merge with all the associated text values for Medline-entry_pub-type before its redundancy removal, but it would not be dependent on other child tables (Figure 3.21b).

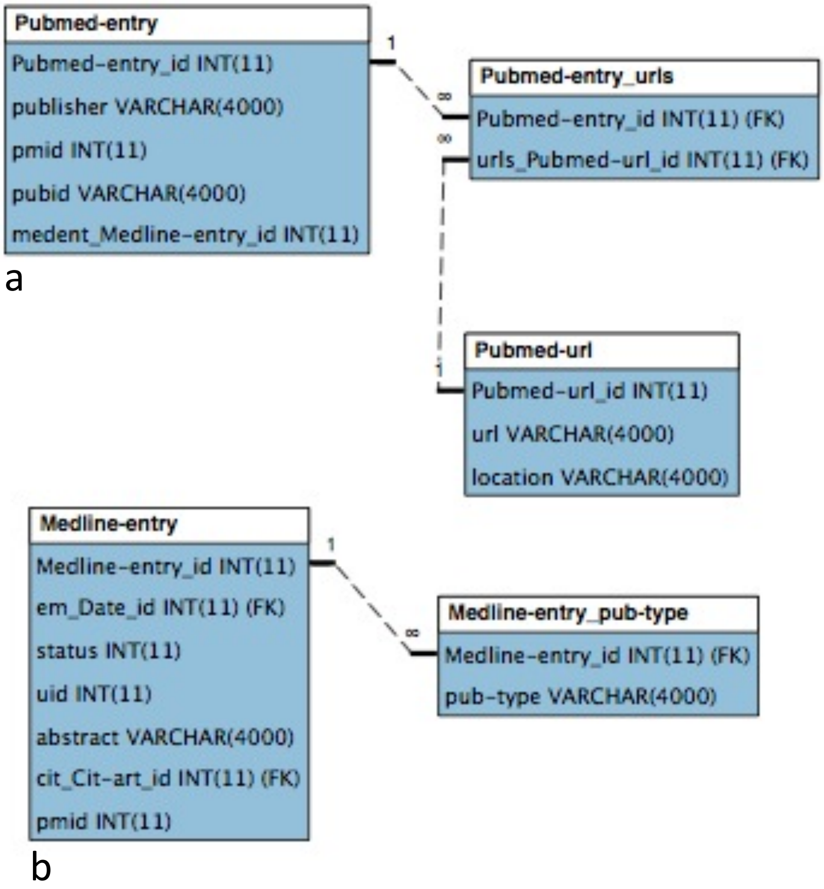


Figure 3.21: Table Relationships a.) Pubmed-entry_urls is the parent of Pubmed-entry and Medline-entry b.) Medline-entry_pub_type is the parent of Medline-entry

There is a great level of dependencies in relational schemas increasing the challenges to prevent conflicts when processes are executed simultaneously. M. C. Er discusses an approach to parallel computation for topological sorting [135]. BioSPIDA applies this algorithm and assigns each table a dependency level. The procedure begins at tables with no children and they are assigned a level 1. Parents of tables with level 1 are assigned level 2. This process continues incrementing the level for each parent table until root tables are reached that have no parents. A table can be reached by multiple paths. In this case, a table is assigned the maximum dependency level of all possible paths. This algorithm can be observed in Figure 3.22. After completion, each table has a unique dependency level. This facilitates BioSPIDA to parallelize the redundancy removal process and sort tables simultaneously. All tables at dependency level 1 are sorted and updated in parallel processes. After all processes are completed, the procedure can then move to the tables at the next dependency level. This greatly expedites redundancy removal, because each operation is no longer restricted to occur sequentially.

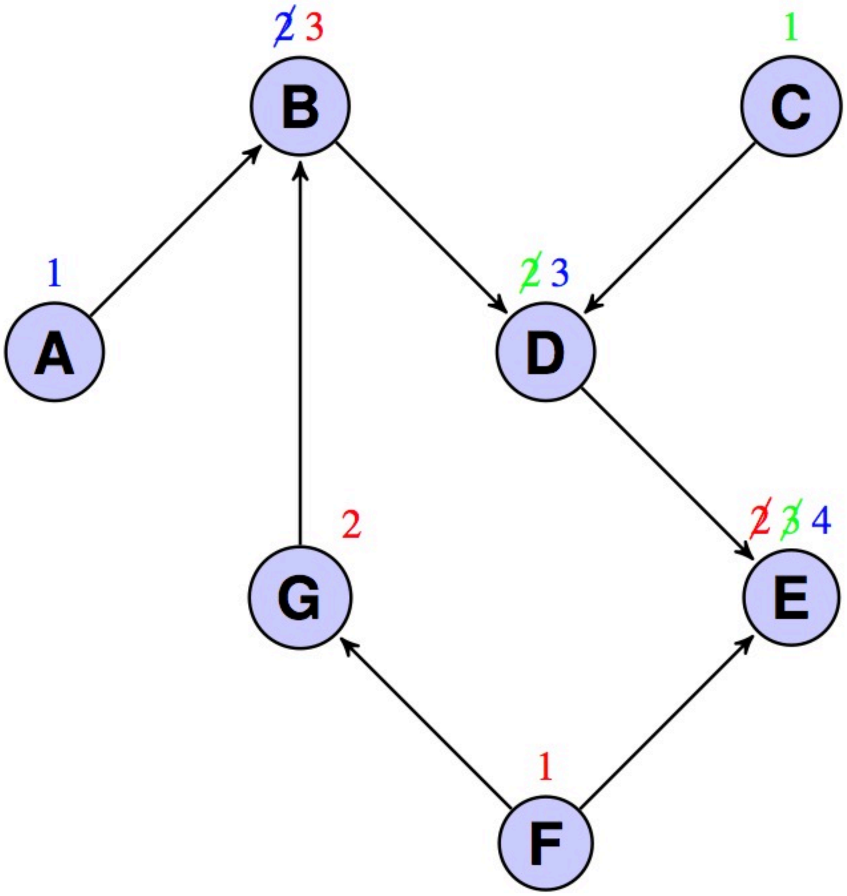


Figure 3.22: Digraph that illustrates ongoing revisions of dependency level for each node. (blue) Dependency level for paths that begin at node A. (green) Dependency level for paths that begin at node C. (red) Dependency level for paths that begin at node F.

Redundancy removal does not only shrink the system storage requirements. It also significantly decreases the database load time. It would be counterintuitive to load billions of records if only half would remain in the final database. After a child table removes its duplicate records, parent tables must be updated. To update a table, its data must be located in the database. This would necessitate loading all initial data and removing duplicates after the records are updated. It is a slow process to sort and delete billions of records within a database. Instead, tables can be updated without loading the data using the unix awk and sort utilities. By utilizing Unix sort and awk utilities, BioSPIDA is able to remove all redundancy from the tables without initially loading any extraneous data. Unix sort performs considerably faster than sorting billions of records within a database and records can be updated before loading any data. Also, sorting outside the database facilitates a second layer of parallelization. Tables at the same depth level can be sorted and updated simultaneously since they do not depend on one another. However, the sorting of an individual table can be processed in parallel as well. If tables grow to sizes larger than 1 Gb, they are split into smaller files and sorted on separate nodes. After sorting is complete of all split files, they are then sort merged by unix and combined into a single ordered table. This process is illustrated in Figure 3.23. Figure 3.23 depicts the sorting, extracting unique rows, and updating steps as each table enters redundancy removal. It displays the second layer of parallelization when larger tables are split and sorted with multiple nodes. The first layer of parallelization is not shown in the figure, because it has already been identified that tables can enter redundancy removal simultaneously if all dependencies are met.

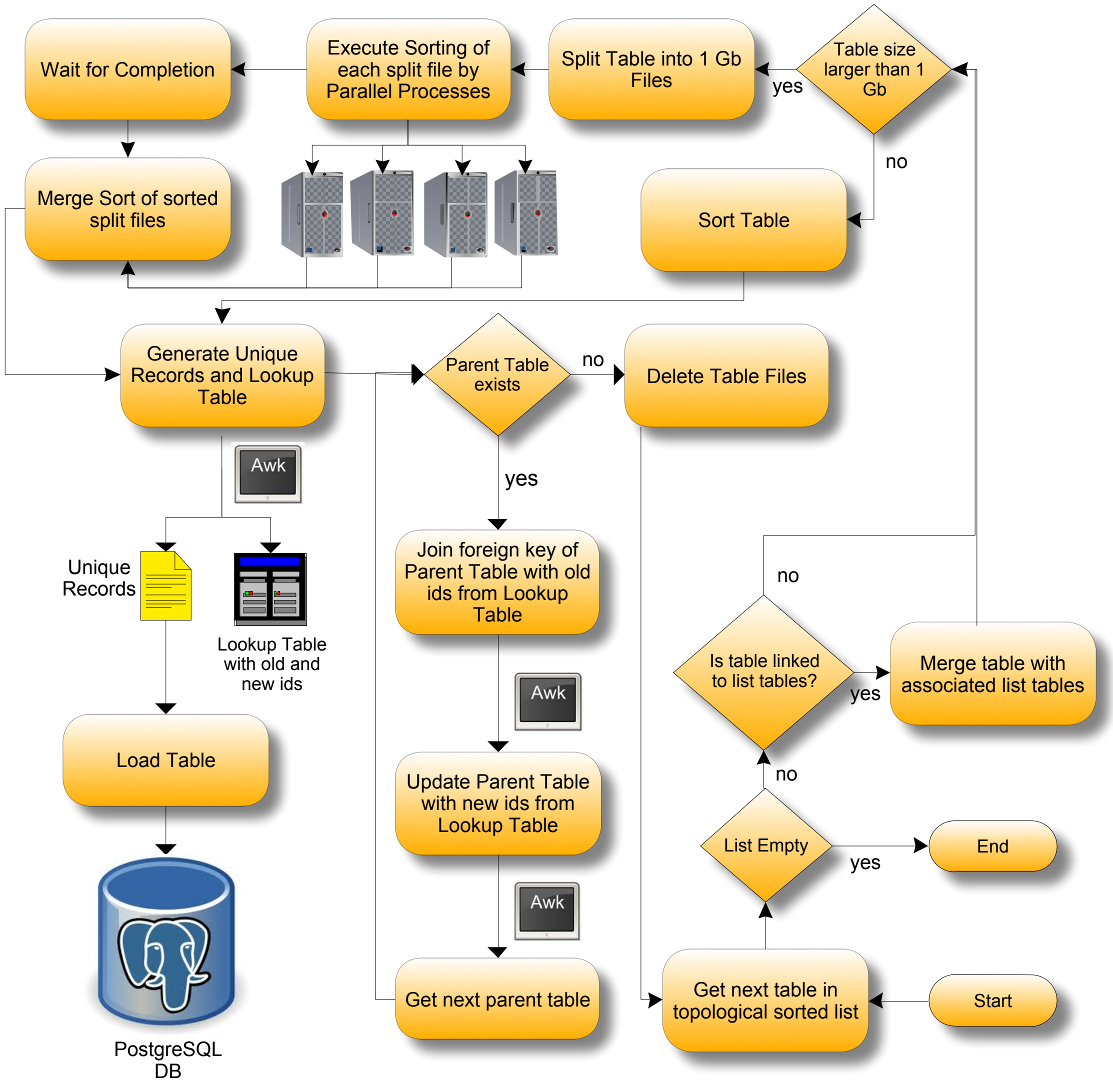

Wait for Completion
Execute Sorting of each split file by Parallel Processes
Split Table into 1 Gb Files
yes
Table size larger than 1 Gb
no
Merge Sort of sorted split files
Sort Table
Generate Unique Records and Lookup Table
Parent Table exists
no
Delete Table Files
Awk
yes
Unique Records
Lookup Table with old and new ids
Join foreign key of Parent Table with old ids from Lookup Table
no
Is table linked to list tables?
yes
Merge table with associated list tables
Awk
Load Table
Update Parent Table with new ids from Lookup Table
no
List Empty
yes
End
Awk
Get next parent table
Get next table in topological sorted list
Start
PostgreSQL DB

Figure 3.23: Redundancy removal process for sorting, removing records, and updating parent tables.

3.4 Data Sources

NCBI provides numerous biological databases under a wide variety of domains, such as sequences, literature, genes, domains, and 3-dimensional structures. BioSPIDA 1.0 [202] was able to parse and convert all NCBI databases developed with an ASN.1 schema. However, ASN.1 is not the only data model used by NCBI along with the many other available biological resources. The 2014 Nucleic Acids Research Database Issue lists a total 1,552 molecular biology databases [147] divided into 14 categories and 41 subcategories. ASN.1 predates XML, and in the past years many in the bioinformatics community have turned to XML to meet the needs for open data exchange in the biological sciences [68]. ASN.1 has its advantages, because it contains a compact binary encoding which allows it to store data in one sixth the size of XML. ASN.1 contains a flexible formal structure while XML defines three different classes of data (ENTITY, ELEMENT, ATTRIBUTE) with arcane and subtle differences [367]. However, the largest advantage of XML is the amount of programmers with a working knowledge and its support by volumes of current software. As a result, BioSPIDA 1.1 has been extended to support the parsing of DTD (Document Type Definition) and (XML Schema Definition) XSD files. Due to many analogous features, support of XSD and DTD schema files could be accomplished without substantial revisions. A CHOICE data type in ASN.1 is represented by xs:choice in XSD, and SEQUENCE is represented by xs:sequence. However, slight modifications were required to handle inherited groups and the subtle differences between xs:element and xs:attribute data types. XML DTD and XML Schema are both backed by W3C as standardized representations for XML documents. XML Schema was developed to replace DTD due to its enhanced features, such as inheritance and user-defined data types [285]. However, many biological resources still use DTD without yet reaching complete adoption of XSD.

A comprehensive listing of available biological resources with associated data types, download addresses, and schema definition files is provided in Table 3.1. The datasets for CDD, Refseq, MMDb, and Entrez Gene can be downloaded in binary ASN.1[10] format by ftp. Pubmed can be downloaded in XML format, but only after

[10] http://www.ncbi.nlm.nih.gov/data_specs/asn/NCBI_all.asn

a license agreement[11]. Records for Pubmed along with Entrez Gene, CDD, Refseq, and MMDb can be retrieved in ASN.1 format by using the NCBI E-utilities [441]. EFetch[12] can retrieve responses for a list of identifiers, but EPost[13] is recommended for larger bulk requests to prevent long URL requests. For BioSPIDA, EPost was used to iteratively send 10,000 identifiers and retrieve ASN.1 records to build the Pubmed database. Full text literature can be downloaded by ftp from Pubmed Central in XML format. The XML is specified with the Journal Article Tag Suite (JATS) format[14] and the DTD[15] and XSD[16] schema files are provided. BioSystems is an NCBI database that describes biological systems and their participating genes, proteins and small molecules. All files can be downloaded by ftp and the data model is available as a DTD[17] or XSD[18] schema file.

[11] http://www.nlm.nih.gov/databases/license/license.html

[12] eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi

[13] eutils.ncbi.nlm.nih.gov/entrez/eutils/epost.fcgi

[14] http://jats.nlm.nih.gov/index.html

[15] http://jats.nlm.nih.gov/archiving/1.0/JATS-archivearticle1.dtd

[16] http://jats.nlm.nih.gov/archiving/1.0/xsd/JATS-archivearticle1.xsd

[17] ftp://ftp.ncbi.nih.gov/pub/biosystems/schema/NCBI_Systems.dtd

[18] ftp://ftp.ncbi.nih.gov/pub/biosystems/schema/NCBI_Systems.xsd

Table 3.1: Data source access for public biological databases.

Data Source	*Data type*	*URL*	*Schema Model*	*Data Format*
NCBI				
BioSystems	pathways	ftp://ftp.ncbi.nih.gov/pub/biosystems/CURRENT	XSD	XML
CDD	domains	ftp://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd	ASN.1	Binary ASN.1
Entrez Gene	genes	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/	ASN.1	Binary ASN.1
MMDb	structure	ftp://ftp.ncbi.nlm.nih.gov/mmdb/mmdbdata/	ASN.1	Binary ASN.1
Pubmed	literature	http://eutils.ncbi.nlm.nih.gov/entrez/eutils/	ASN.1	ASN.1
Pubmed Central	literature	http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/	JATS	XML
Refseq	sequence	ftp://ftp.ncbi.nlm.nih.gov/refseq/release/complete/	ASN.1	Binary ASN.1
EMBL-EBI				
Expression Atlas	expression	http://www.ebi.ac.uk/gxa/download.html		Delimited Text
IntAct	interaction	http://www.ebi.ac.uk/intact/pages/documentation/downloads.xhtml	PSI MI	XML
GO	ontology	http://geneontology.org/GO.downloads.database.shtml	MySQL	MySQL dump
Interpro	domains	http://www.ebi.ac.uk/interpro/download.html	DTD	XML
Reactome	pathways	http://www.reactome.org/download/	MySQL	MySQL dump
Uniprot	proteins	http://www.uniprot.org/downloads	XSD	XML
Other Sources				
BioGRID	interaction	http://thebiogrid.org/download.php	PSI MI	XML
DIP	interaction	http://dip.doe-mbi.ucla.edu/dip/Download.cgi	PSI MI	XML
HPRD	interaction	http://www.hprd.org/download	PSI MI	XML
KEGG	pathway	http://www.kegg.jp/kegg/download/	KGML	XML
MINT	interaction	http://mint.bio.uniroma2.it/mint/download.do	PSI MI	XML
OMIM	phenotype	http://www.omim.org/downloads		Delimited Text
STRING	interaction	http://string-db.org/newstring_cgi/show_download_page.pl		Delimited Text

The European Bioinformatics Institute (EMBL-EBI) is a research institution supported by over 20 European countries with approximately 85 independent groups[19]. They offer a comprehensive range of freely available biological databases, including proteins, gene expression, pathways, and domains. Uniprot [28] is a database of protein sequences and functional information is downloadable by ftp in XML. Only an XSD[20] schema file is provided. Interpro [225] is a database of families and predicted domains of protein sequences. All individual matched signatures for all proteins in the Uniprot Knowledgebase are downloaded by ftp in XML format. Only a

[19] http://www.ebi.ac.uk/services/all

[20] http://www.uniprot.org/docs/uniprot.xsd

DTD[21] schema file is provided. IntAct is a molecular interaction database derived from literature curation and direct submissions. The data is available in PSI-MI 2.5 format. PSI-MI format was developed by the Proteomics Standards Initiative (PSI) to standardize data exchange for molecular interactions [213]. It is also described with an XML Schema model[22]. Most protein interactions databases provide their datasets by ftp in PSI-MI 2.5 XML format, including HPRD [397], MINT [76], DIP [430], and BioGRID [471]. This gives BioSPIDA 1.1 the ability to automatically convert these databases, since an XSD schema file is used to describe the PSI-MI data model. Other example databases supported by EMBL-EBI are Reactome [111], Gene Ontology (GO) [20], and Expression Atlas [251, 252]. Reactome is a peer-reviewed pathway database containing biological networks and participating nucleic acids, proteins, and other molecules. The database is available as a MySQL dump and no parsing is needed. The Gene Ontology Consortium provides a controlled vocabulary and ontology of terms for describing gene product characteristics. The database is also available as a MySQL dump. The Expression Atlas provides gene expression patterns under many biological conditions and are filterable by criteria, such as developmental state, organism part, environment factor, and cell type. The data for all differential expression analysis results are downloadable in tab delimited format. Other independent resources are Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) [156,235,464,483,514–516], Online Mendelian Inheritance in Man (OMIM) [206], and Kyoto Encyclopedia of Genes and Genomes (KEGG) [248]. STRING is a database of predicted protein interactions derived from literature, co-expression, and genomic context. Data is publicly available as tab delimited files or via license approval for a full database dump. OMIM is a phenotype database that contains information on all known Mendelian disorders on over 12,000 genes. Data

[21] ftp://ftp.ebi.ac.uk/pub/databases/interpro/match_complete.dtd

[22] http://psidev.sourceforge.net/mi/rel25/src/MIF25.xsd

is in tab delimited format and is freely available by ftp by application. KEGG is another comprehensive database for biological pathways as well as for diseases and drugs. Data is downloadable by ftp but only for paid subscribers. KEGG does offer an API[23] for record retrieval and pathways are provided in KEGG Markup Language (KGML). KGML is an XML data type described by a DTD[24] specification file. Biological data centers are maintained by many separate teams under different scientific domains. This review accurately illustrates how quickly complexity can arise, especially when there is an absence of a parent organization that enforces universally accepted data exchange standards.

A data warehouse was built from many available public resources to create a locally integrated system with a diverse set of biological data types. Pathways and annotations were downloaded from the MySQL dumps of Reactome and GO. The MySQL schema and loading files were converted into Postgres and inserted into the relational database. Expression Atlas, STRING, HPRD and OMIM were all directly loaded into Postgres after download of tab delimited files. This required some manual query creation of table names and indexes. After Pubmed and Entrez Gene completed redundancy removal, both databases were created and loaded into Postgres by BioSPIDA. The complete system of genes, pathways, annotations, literature, interaction, and phenotypes can be viewed in Figure 3.24.

[23] http://www.kegg.jp/kegg/rest/

[24] http://www.kegg.jp/kegg/xml/KGML_v0.7.1_.dtd

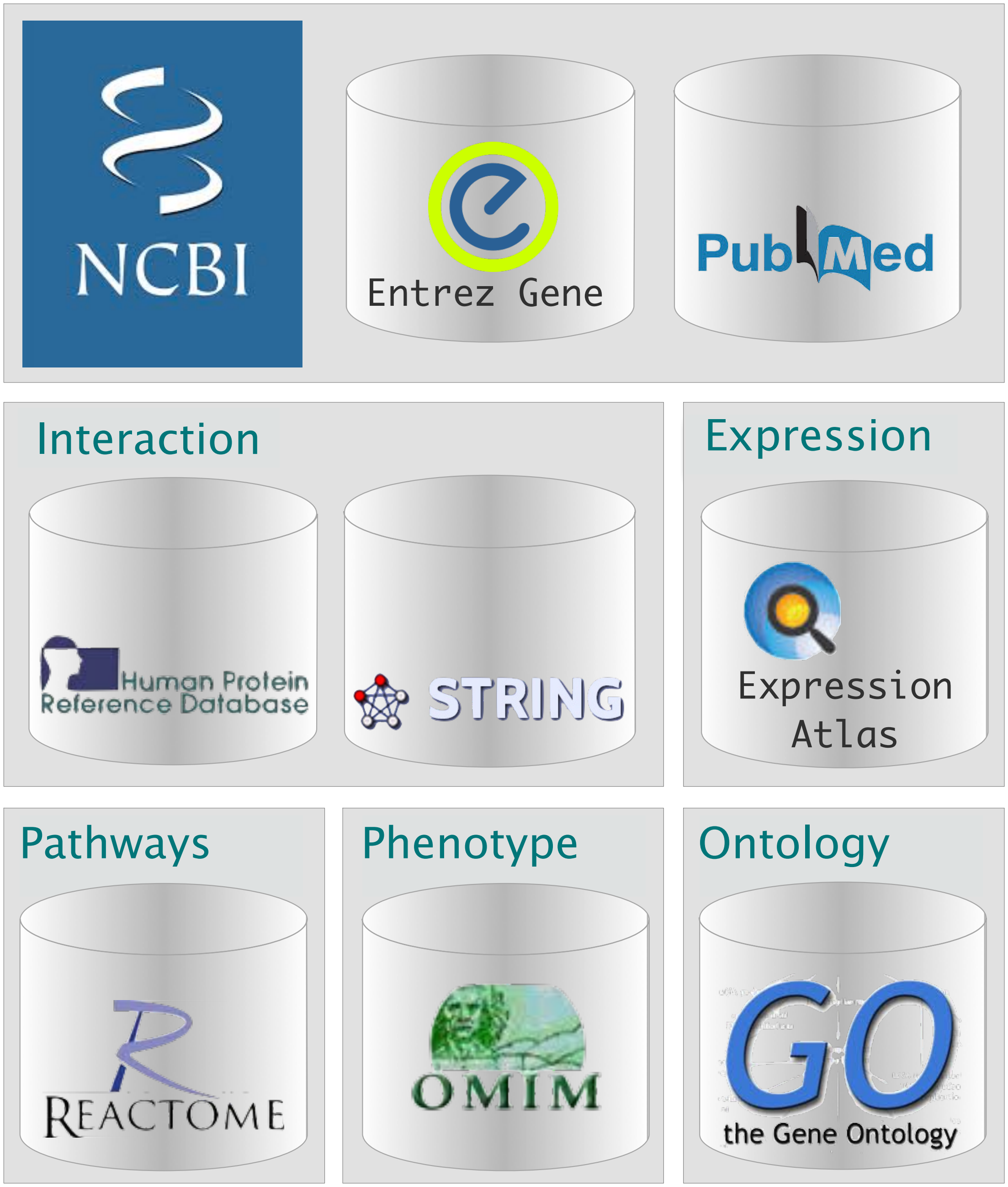


Figure 3.24: Data warehouse for integrated biological resources.

3.5 Results

BioSPIDA 1.1 was tested against six separate NCBI databases: Entrez Gene, Pubmed, CDD, Refseq, MMDB, and Biosystems. The fully extracted XML Data for each NCBI database could grow to considerable storage sizes, such as 302 Gb for Entrez Gene, 208 Gb for Pubmed, and 2.4 Tb for Refseq. Fortunately, XML extraction is only required once per record and could be deleted after conversion. The XML for each database is translated into tab delimited database loading files. Sequentially, this process would have required significant workload to fully convert an entire NCBI database. Parsing all of Pubmed would have totaled 14.5 days and Entrez Gene and Refseq would have completed in 14.5 days and 179.3 days. These durations are not realistic for researchers that wish to have a locally integrated database of NCBI resources. Each database was parsed in parallel with 16 separate nodes. The parallel runtime reduced to 52.6 hours, 28.7 hours, and 12.3 days for Entrez Gene, Pubmed, and Refseq. In converting the XML data to raw loading files, there was also a substantial size reduction. For Entrez Gene, Pubmed, and Refseq, data sizes decreased to 83.7 Gb (28%) , 62.4 Gb (30%), and 669.12 Gb (28%). For MMDb, the converted data size was 140.74 Gb (18%) from 730.79 Gb, most likely due to the high XML tag usage when describing 3-dimensional structures. In translating NCBI ASN.1 schemas to a relational database, a significant amount of tables were created to accurately describe each data model. 128 tables were generated for Entrez Gene and 499 for CDD. Manual creation of these high numbers of tables with accompanying field descriptions would require a significant undertaking, and the relational translator eliminates these arduous tasks from development of the local data warehouse. Additionally, a smaller subset of Homo Sapien records for Entrez Gene was tested in the event that an entire dataset is not necessary for research. All runtimes and storage sizes are viewable for each NCBI database in Table 3.2.

Table 3.2: Parallel runtime and data size output for XML Parsing of NCBI Databases.

Database	*XML Data*	*Raw Data*	*Records*	*Tables*	*Parallel runtime*	*Sequential runtime*
Biosystems	63.68 Gb	32.3 Gb	579,908	353	6.28 h	6.24 d
CDD	61.50 Gb	21.31 Gb	48,280	499	7.61 h	4.7 d
Entrez Gene	302.01 Gb	83.70 Gb	12,234,134	128	52.67 h	25.6 d
Entrez Gene†	8.48 Gb	3.28 Gb	194,720	128	8.83 h	11.74 h
MMDb	730.79 Gb	140.76 Gb	220,825	224	23.02 h	15.2 d
Pubmed	207.65 Gb	62.4 Gb	21,212,193	43	28.7 h	14.5 d
Refseq	2.37 Tb	669.12 Gb	10,071	339	12.3 d	179.3 d

† Homo Sapien subset
days(d), hours (h)

Redundancy removal was conducted on the Entrez Gene, Pubmed, and Homo Sapien subset of Entrez Gene to further compare workload and storage space sizes. The total amount of rows reduced almost 3:1 for each database. The final amount of records were 1.3 billion (37%) , 645 million (26%), and 34 million (34%) in Entrez Gene, Pubmed, and Entrez Gene (Homo Sapien). The storage size reductions were roughly 2:1 after executing redundancy removal on each database. The total sizes were 41.44 Gb (50%) , 62.4 Gb (50%), 995.7 Mb (30%) for Entrez Gene, Pubmed and Entrez Gene (Homo Sapien). However, the data sizes decreased even more sharply when compared to the original extracted XML files at 14%, 15%, 11% for Entrez Gene, Pubmed, and Entrez Gene (Homo Sapien). Although, the record sizes did not decrease at the same ratio as the storage sizes, this is still significant because smaller record volumes lead to smaller indexes and faster query execution. Entrez Gene (Homo Sapien) may have had a greater reduction, because the focus on one species dataset could lead to higher record redundancies. The results can be viewed in Table 3.3.

Table 3.3: Redundancy removal data size output.

Database	*XML Data*	*Raw Data*	*Reduced Data*	*Rows*	*Reduced Rows*
Entrez Gene	302.01 Gb	83.70 Gb	41.44 Gb	3,576,481,900	1,317,650,325
Entrez Gene†	8.48 Gb	3.28 Gb	995.7 Mb	102,343,424	34,361,458
Pubmed	207.65 Gb	62.4 Gb	31.06 Gb	2,458,574,541	645,501,015

† Homo Sapien subset

Runtimes for redundancy removal on Pubmed, Entrez Gene and the Homo Sapien subset were examined to identify differences in execution speeds when focusing on a smaller integration project. Redundancy removal was executed in parallel, and a separate node was allocated for a table that requested sorting, merging, and updating operations. A maximum of 8 nodes could be allocated simultaneously. However, it was rare for more than 8 nodes to be requested, because most tables could not start redundancy removal until all of their dependent tables had completed. As stated in Section 3.3.6, there were two layers of parallelization. Tables that were not dependent on each other could both execute redundancy removal simultaneously. The second layer of parallelization is if a very large table needs to be sorted. Tables larger than 1 Gb would be split, and smaller files were sorted in parallel and then merged. The total redundancy removal process for the entire Entrez Gene and Pubmed datasets were 61.6 hours and 20.9 hours. This included merging tables with associated list tables, sorting tables, extracting unique rows and lookup tables, and updating parent tables with new ids. If there were no layers of parallelization, redundancy removal on tables would occur sequentially, and sorting large tables could not split files and spawn child processes. A total sequential process would have taken 300.3 hours and 198.3 hours for all of Entrez Gene and Pubmed. The two layers of parallelization created a 5:1 improvement for Entrez Gene and almost a 10:1 ratio for Pubmed. The ratio is not significantly higher because many operations cannot occur simultaneously, due to table dependencies. Redundancy removal on Entrez Gene spends a total of 61.2 hours waiting time for operation completion. However, the entire process does not idle during waiting time. Waiting time accumulates when at least one node is ready to start redundancy removal on a table, but the process can not initiate because all dependencies are not yet met. For the Homo Sapien subset, there is only one layer of parallelization, because no table ever exceeds 1 Gb. Sequentially, the entire operation would have completed in 4.49 hours. Parallelization reduces execution time to 3.91 hours and total waiting time is 3.84 hours. Again, this is the total waiting time accumulated for all nodes and the overall process is not idle. However, the speed increase is not as significant as in the entire dataset for Entrez Gene. This is most likely due to most operations on the smaller Homo Sapien subset are more accelerated due to smaller data sizes, as evident in a duration of only 24 minutes for merging, extracting, and updating. When so many operations can be executed quickly, the longest operation will carry the majority of the workload and parallelizing smaller operations will not significantly reduce the runtime. The runtimes for all redundancy removal operations can be viewed in Table 3.4.

Table 3.4: Runtimes for redundancy removal and loading table data.

id	Action item	Pubmed	Entrez Gene	Entrez Gene†
1	Merging tables with associated list tables	2.59 h	5.58 h	9.04 m
2	Sorting Tables (with parallelization)	77.9 h	118.4 h	4.09 h
	Sorting Tables (no parallelization)	113.5 h	281.8 h	
3	Extracting unique rows and id lookup table	2.32 h	7.96 h	8.34 m
4	Updating parent table loading files with new ids	2.03 h	5.01 h	6.61 m
5	Waiting for dependent tables to finish sorting or updating	20.9 h	61.2 h	3.84 h
	Total runtime for redundancy removal (parallel)	20.9 h	61.6 h	3.91 h
	Total runtime for redundancy removal (sequential)	198.34 h	300.3 h	4.49 h
	Total runtime for loading Postgres database	13.5 h	16.8 h	15.06 m

† Homo Sapien subset
hours (h), minutes (m)

The total data size for all databases was 676.4 Gb, and in some cases the storage sizes for indexes grew larger than row data. This is especially true for databases that required many joins but without as much raw textual content, such as Reactome, GO, and Entrez Gene. All storage sizes for row and index data can be viewed in Table 3.5. Entrez Gene and Pubmed both total over 250 Gb, but each are a local full conversion of an entire NCBI database containing 12 million genes and 24 million biomedical publications. Both databases are also results of redundancy removal where an average of 70% of records were eliminated. Without redundancy removal, more substantial storage space would haven been required, and data reductions made a locally integrated warehouse of entire NCBI databases much more feasible.

Table 3.5: Row and index size for integrated data warehouse of biological databases.

Database	*Index size*	*Row size*	*Total size*
Entrez Gene	101 Gb	66 Gb	167 Gb
Entrez Gene[†]	2.70 Gb	1.67 Gb	4.36 Gb
Expression Atlas	27 Gb	39 Gb	66 Gb
GO	156 Gb	49 Gb	206 Gb
HPRD	28 Mb	91 Mb	120 Mb
OMIM	177 Mb	277 Mb	453 Mb
Pubmed	50 Gb	49 Gb	99 Gb
Reactome	790 Mb	725 Mb	1515 Mb
STRING	68 Gb	163 Gb	231 Gb

[†] Homo Sapien subset

The final integrated data warehouse contains Entrez Gene, Expression Atlas, GO, HPRD, OMIM, Reactome, and STRING. In one local relational database, information can be extracted and linked pertinent to gene expression, functional annotation, protein interaction, metabolic pathways, biomedical literature and disease. This facilitates richer and more customizable queries that are not possible with a navigational website provided by many biological resource sites. In Entrez Gene, genes are associated with biochemical details including metabolic pathways, protein interactions, and gene ontology. However, most annotations in Entrez Gene are stored in Gene-commentary fields that can become deeply nested and recursively linked to itself. This can lead to SQL queries that require a significant amount of table joins. The query execution time to extract 3.67 million protein interactions from all genes in the entire Entrez Gene dataset was 41.5 minutes. However, only one execution is required, because the results can be stored as another table or a materialized view and used for future querying and filtering. This was also a more complicated example that contained three levels of Gene-commentary fields nested in each other. There are many benefits to integrating a subset of Entrez Gene Homo Sapien data as well. In this smaller dataset, 305,597 protein interactions were extracted for 18,089 genes in 8.28 seconds. 27,081 pathways were found for 7,123 genes in 2.81 seconds. 197,010 gene ontology annotations for 14,885 genes were found in 5.81 seconds. Additionally, in converting Biosystems XML data to relational tables, genes could be associated to many more metabolic pathways than identified in Entrez Gene. In Entrez Gene, an average of 3.81 pathways were linked to each gene. By integrating NCBI Biosystems, 10,288 genes contained a total of 134,333 pathways with an average 13.05 pathways per gene. This is a significant advantage of locally integrated data. As an example, to receive details for only 1,000 records from Biosystems through the Web Service BioPython [75, 101] took a total of 9.53 minutes. With the usage of indexes and a locally integrated relational database, many other potential linkages can be quickly queried across biological resources, such as between diseases and relevant publications or between overexpressed genes and known protein interactions. Using Expression Atlas, 588 human genes were identified to be over expressed in the heart. The associated publications, abstracts, and mesh terms for these genes were retrieved in 0.74 seconds with a query linking Expression Atlas, Entrez Gene and the Pubmed database. The execution was on the complete Entrezgene dataset, so similar queries can be rapidly processed for other organisms.

3.6 Discussion

It is clear that efficient memory management, parallelization, and redundancy removal provide a powerful platform to integrate biological sources. Researchers require an environment that can quickly answer complex questions without manually visiting many locations to find rich associations. Data structures for the biomedical sciences are updated continuously and investigators are either dependent on the service level of outside centers or are faced with heavy programmatic challenges of creating a local repository for heterogeneous sources. BioSPIDA automatically generates NCBI relational databases and removes the necessity of developing custom source code. Setup time and storage requirements are reduced by the newest feature enhancements. This facilitates investigators to take advantage of the benefits provided by data warehouses while minimizing their high overhead shortcomings.

Many further studies can take advantage of a locally integrated data warehouse with diverse biological sources that are easily linkable. The relational database contains many data sources that facilitate more powerful queries that can expand filters by multiple criteria. Records can be explored at a deeper level of granularity. The Reactome database provides more than just the name of a biochemical pathway. Each Reactome event contains reactions with involved substrates, enzymes and molecular products. The Gene Ontology Database includes genome annotations for molecular functions, biological processes and cellular components all organized hierarchically. This allows a more sophisticated comparison between genes where identifying common functionality is not only limited to shared annotations. With an ontological graph structure, semantic similarity can be computed between genes by the shortest paths between their annotations. All associated references for a set of genes can be investigated and a natural language processing study could be conducted to identify the most prevalent keywords. The NCBI ELink[25] utility can also be utilized to retrieve linked records to other NCBI databases from a set of genes of interest. This would negate the need to fully integrate an entire NCBI database if unnecessary.

There are other possible enhancements that can help further improve the performance of the parsing and integration process. For redundancy removal, the most intensive operation is the sorting process. It can be unnecessary to sort and search for duplicate rows if some fields have very long values. It is unlikely if a field value is over

[25] http://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.ELink

1,000 characters that there will be another row with its exact duplicate. Entrez Gene contains four tables that contain fields that reach lengths longer than 1,000 characters. Each of these tables were longest in the sorting process. Redundancy removal cannot move to the next dependency level until all tables at the current level are sorted. Therefore, the total runtime of sorting tables at each level can only execute as quickly as the time to sort the largest table. By ignoring rows with longer field values, it can significantly speed the sort time of large tables without possibly large changes in data reduction. It may also not be necessary to create an index on every table. Many tables may not currently be utilized for a current study, and it could save substantial storage space size by reducing indexes. Potential enhancements will continue to be explored to help develop a compelling system that seamlessly integrates heterogeneous biological sources.

CHAPTER IV

PRIORITIZING CANDIDATE DISEASE GENES

4.1 Introduction

The current breakthroughs of genome sequencing and high-throughput experimental techniques have led to the development of large amounts of quantified data on biological systems. However, we are still far from a complete understanding behind the intricate molecular networks involved in human diseases. Many common diseases are hereditarily complex [181] and are influenced by the affects of many genes, and even simple Mendelian traits can be subsequently altered by modifier genes [52]. Dissecting genetic architectures has profound implications to the treatment of human disease, but it remains extraordinarily difficult to identify causal genes [51]. Due to the pleiotropy of genes, genetic heterogeneity, and limited number of cases, it is a challenging task to fully link the relationship between genetic variation and clinical phenotypes [179]. Traditionally, discovery of causal genes has been accomplished with statistical methods, such as linkage analysis and positional cloning [52]. However, the limitations of these approaches have restricted the identification of genetic associations mainly to Mendelian disease phenotypes [389].

Genome-wide association studies have proven more successful in determining candidates for complex forms of disease, but genome-wide techniques often select loci containing hundreds of potential candidates. Manually identifying the most promising candidates by reviewing available biological resources would be a daunting task, due to the large variety of public databases [343]. No single source of data contains all relevant relations, and verification of genetic associations would require visiting sites pertaining to functional annotation (KEGG [248], Reactome [111], Gene Ontology

(GO) [20], HPRD [397]), expression data (Expression Atlas [251, 252], GEO [33]), and sequence data (Ensembl [150], Entrez Gene [314]). This overwhelming wealth of biological data has posed extra challenges to scientists as opposed to original expectations. Currently, there is a high demand for analytical tools that can accurately integrate, mine, and prioritize relevant information for determining gene candidates in disease [30, 247]. A single study focusing on coexpression would miss the effects of post-translation modification, and a study only targeting protein-protein interaction would disregard transcription regulation. Merging diverse types strengthens the current known evidence and provides a much broader coverage than any single data source could accomplish by itself [343].

4.2 Related Work

Most existing computational methods that prioritize candidate disease genes are based on the guilt-by-association principle [184, 463] where associations are discovered by functional or topological similarities to known disease genes. Previous predictive studies have taken advantage of available resources, such as gene ontology annotations [159,380], published literature [238], protein interactions [307,380,383] and gene expression [307, 492]. However, presently, disease-gene prediction with application of only one type of evidence has become more rare. Multiple sources can provide complementary evidence and as a whole, a larger picture can be delivered regarding the inner workings of molecular mechanisms. Zhang et al. [545] focus on protein interaction networks (PPIs), but still integrate multiple data sources by analyzing five separate databases. Using a Bayesian Regression approach, the integration of multiple protein interaction networks performed higher than any other network alone. Interestingly, it was found by Köhler et al. [271] that global measures perform better than local measures when prioritizing with PPIs. ENDEAVOUR was one of the earlier prioritization tools that integrate multiple data sources [4, 495]. The ranks from each data source are combined into an overall score, but it has the disadvantage that each data source must have its own metric.

More contemporary prioritization approaches use a unified framework where new data sources may be added without developing a new methodology. Li and Patra [293] incorporate gene ontology and a PPI to create a discounted rating system where the scores from protein interaction, molecular function, cellular component, and biological process are effectively weighted and combined. Linghu et al. [299] build an evidence-weighted functional linkage network with a Naive Bayes classifier based on 16 separate genomic features, including protein domains, literature co-occurrence, ontology, and protein interaction. DIR, developed by Chen et al. [88], utilize multiple data sources by defining networks with diffusion kernels. Diffusion kernels [274] are a global measure of similarity and have outperformed metrics that only depend on immediate neighbors. DIR uses these measures to rank candidate genes and only the most informative network contributes to the overall score. The current study in this work is the **W**eighted **IN**tegration of **G**ene **S**ources for **P**rioritization **R**anking of **E**ach **A**nnotation to **D**isease (WINGSPREAD). WINGSPREAD also uses multiple data sources and diffusion kernels to prioritize disease-gene associations. However, advanced data integration strategies and semantic similarity measures are utilized to develop edge weights for each network. Furthermore, multiple networks can influence the final score and are appropriately discounted based on the network informativeness.

4.3 Diffusion kernels

Kernels can efficiently handle complex data types by appropriately defining a similarity matrix between input data objects. A kernel matrix must satisfy the mathematical requirements of being symmetric and positive definite [143]. Positive definiteness is defined that for any choice of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathcal{X}$ and coefficients $c_1, c_2, \ldots, c_m \in \mathbb{R}$,

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \tag{4.1}$$

Kernels can be used for graph similarity metrics, but it is not clear how to construct a positive definite matrix. Random walks over graphs start at a given vertex z_0 and for each timestep $t = 1, 2, \ldots, T$ choose a new vertex z_t from the neighbors of z_{t-1}. This process can be represented by a normalized adjacency matrix

$$Q_{ij} = \begin{cases} 1/\gamma_{ij} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where γ_{ij} is the degree of vertex j. The matrix Q is treated as a transition matrix for the random walker at each vertex. The probability of finding a random walker at node j at time t is $p_{t+1} = Q\,p_t$. This relation can be applied recursively by raising the normalized adjacency matrix to the number of timesteps T:

$$P_T = Q^T \tag{4.3}$$

However, P_T is not a suitable choice for a kernel, since there is no clear choice for T, particular choices of T may make certain vertices unreachable, and P_T is generally not positive definite [447].

The exponential of a square matrix H is defined as:

$$e^{\beta H} = \lim_{n \to \infty} \left(1 + \frac{\beta H}{n}\right)^n \tag{4.4}$$

It can be shown that an infinitely divisible kernel K can be expressed in exponential form [274] suggesting the usage of kernels for continuous random walks that satisfy the equation:

$$K = e^{\beta H} \tag{4.5}$$

This results in the following differential equation:

$$\frac{d}{d\beta} K_\beta = H K_\beta \tag{4.6}$$

The *graph Laplacian* is defined as:

$$L_{ij} = \begin{cases} -1 & \text{if } i \sim j \\ \gamma_{ij} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

The *negative graph Laplacian* is an ideal candidate for H in equation 4.5 because it satisfies the differential equation 4.6:

$$\begin{aligned} \frac{dk_i}{dt} &= -\sum_j A_{ij}(k_i - k_j) \\ &= -k_i \sum_j A_{ij} + \sum_j A_{ij} k_j \\ &= -k_i\, deg(v_i) + \sum_j A_{ij} k_j \\ &= -\sum_j (\delta_{ij}\, deg(v_i) - A_{ij})\, k_j \\ &= -\sum_j (L_{ij})\, k_j \\ \frac{dK}{dt} &= -LK \end{aligned}$$

Usage of the *negative graph laplacian* as H in the differential equation 4.6 is called the heat equation and the resulting kernels are known as *diffusion kernels* or *heat kernels*.

We also can prove that diffusion kernels are positive definite, since any even power of a symmetric matrix is also positive definite:

$$e^{\beta H} = \lim_{n \to \infty} \left(1 + \frac{\beta H}{n}\right)^n = \lim_{2n \to \infty} \left(1 + \frac{\beta H}{2n}\right)^{2n} \tag{4.8}$$

Matrix exponentiation shares many properties with ordinary exponentiation [186]. This includes the power series expansion:

$$e^{\beta H} = I + H + \frac{1}{2!} H^2 + \frac{1}{3!} H^3 + \ldots \tag{4.9}$$

Using these properties, the matrix exponentiation can be computed in a finite amount of time by calculating the normalized eigenvectors of $v_1, v_2, \ldots v_n$ and corresponding eigenvalues of $\lambda_1, \lambda_2, \ldots \lambda_n$ of the negative graph Laplacian L:

$$L^s = \left(\sum_{i=1}^{n} v_i \lambda_i v_i^T \right)^s = \sum_{i=1}^{n} v_i \lambda_i^s v_i^T \tag{4.10}$$

$$e^{\beta L} = I + \left(\sum_{i=1}^{n} v_i \beta \lambda_i v_i^T \right) + \left(\sum_{i=1}^{n} v_i \frac{(\beta \lambda_i)^2}{2} v_i^T \right) + \ldots = \sum_{i=1}^{n} v_i e^{\beta \lambda_i} v_i^T \tag{4.11}$$

In WINGSPREAD, the negative graph Laplacian is calculated for similarity matrices from six separate data sources. The normalized eigenvectors is then computed for each Laplacian, and the β can be tuned by the summation operation in equation 4.11.

4.4 Methods

4.4.1 Data sources

4.4.1.1 Protein interaction

Disease genes are often highly connected and participate in the same interaction networks [170, 480]. Differences in the network properties have been used to build many candidate gene prioritization methods [4, 84, 85, 88, 157, 174, 177, 271, 279, 287, 296, 342, 404, 451, 545]. As shown by Zhang et al. [545], utilizing several protein interaction networks can also improve accuracy. In the current study, three protein interaction networks were downloaded from Human Protein Reference Database (HPRD) [397], the Biological General Repository for Interaction Datasets (BioGRID) [471], and Biomolecular Interaction Network Database (BIND) [27]. HPRD contains human protein-protein interactions, and BioGRID contains protein and genetic interactions. BIND covers high-throughput and manually curated molecular interactions. All redundant links were removed, and an integrated interaction network was created for all Homo Sapien genes.

4.4.1.2 Gene Ontology

The Gene Ontology system is widely used as a standard for functionally annotating large sets of genes [20]. It consists of three structured ontologies describing molecular function (MF), biological process (BP), and cellular component (CC). Goh et al. [184] showed that genes associated with the same disorder have significantly higher (GO) homogeneity. This knowledge can be used to identify uncharacterized disease genes and has been utilized frequently in previous integrative prioritization approaches [3, 4, 85, 86, 157, 174, 271, 293, 296, 299, 342, 366, 381, 404, 451, 497].

With semantic similarity measures, the functional relationships between gene products can be calculated even if protein pairs do not share any annotations. Concepts are represented in a hierarchical ontological tree, and semantic similarity can be calculated by edge-based or node-based measures. Edge-based methods rely primarily on counting the number of edges linking GO terms in the ontological directed acyclic graph (DAG). Node-based methods rely on comparing the individual properties of compared terms. The information content (IC) is a common property measure for ontological terms which returns the level of specificity or informativeness. IC is defined as the negative log likelihood of a term c:

$$IC(c) = -log\, p(c) \tag{4.12}$$

where p(c) is the probability of occurrence of term c in a specific corpus. Node-based methods have several advantages, since they are not as sensitive to the intrinsic structure of DAG and terms at the same level are not treated equivalently. However, in node-based methods the IC is calculated based on occurrences in an external corpus that is constantly changing. They also can be biased with a preference for well-annotated terms. See Section 2.3 for a more complete survey of ontological similarity measures. In biological problems, no method is clearly preferred over all others and the most most suitable method usually depends on the context [197, 385].

Wu et al. [530] developed an edge- and node-based hybrid method, known as Hybrid Relative Specificity Similarity (HRSS), that consistently outperformed competing methods in determining functional similarity between gene products. It was compared against several edge-based [176,232,531] and node-based [240,297,384,411] methods by inferring physical protein-protein interactions and estimating the functional conservation of orthologs. HRSS is calculated based on the scores of three components, α, β, and γ. α measures the specificity of the terms lowest common ancestor. β measures the generality of the compared terms. γ measures the local distance between terms and their lowest common ancestor. All of these terms are calculated according to information content measures:

$$dist_{IC}(u,v) = IC(v) - IC(u) = log\, p(u) - log\, p(v) \tag{4.13}$$

$$\alpha_{IC} = dist_{IC}(root, MICA) = -log\, p(MICA) \tag{4.14}$$

$$\beta_{IC} = \frac{dist_{IC}(term_i, MIL_i) + dist_{IC}(term_j, MIL_j)}{2} \tag{4.15}$$

$$\gamma = dist(MICA, term_i) + dist(MICA, term_j) \tag{4.16}$$

where MICA is the most informative common ancestor of $term_i$ and $term_j$, MIL_i and MIL_j are the most informative leaf nodes of $term_i$ and $term_j$. The final equation for HRSS is the following:

$$HRSS(term_i, term_j) = \frac{1}{1+\gamma} \cdot \frac{\alpha_{IC}}{\alpha_{IC} + \beta_{IC}} \tag{4.17}$$

The best-match average (BMA) strategy finds the best semantic similarity value for each term and calculates the average:

$$HRSS_{BMA}(P,Q) = \frac{\sum_{tp_i \in TP} HRSS(tp_i, TQ) + \sum_{tq_j \in TQ} HRSS(tq_j, TP)}{\mid TP \mid + \mid TQ \mid} \tag{4.18}$$

where P and Q are gene products, TP and TQ are the sets of all GO terms assigned to P and Q, and $HRSS(u_i, V) = \max_{v_j \in V}\{HRSS(u_i, v_j)\}$.

All Homo Sapien genes with annotated GO terms were retrieved from the NCBI Entrez Gene [313] database. The HRSS similarity score was calculated for all BP and MF GO terms for all pairs of Homo Sapien genes. Similar to the method for the funSim score by Schlicker et al. [445], the final semantic similarity measure was determined by squaring the MF score and BP score:

$$GOSemSim = \frac{1}{2}\left[\left(\frac{BPscore}{max(BPscore)}\right)^2 + \left(\frac{MFscore}{max(MFscore)}\right)^2\right] \tag{4.19}$$

where BPscore and MFscore is the HRSS similarity score between gene products BP and MF GO terms.

4.4.1.3 Domains

Candidate genes may share similar functions to disease genes [242], and these functions may be due to common protein domains. Protein domains are functional units of proteins, and previous methods have incorporated domain knowledge to improve candidate gene prioritization [3, 4, 165, 177, 299, 497]. It was found in many of these studies that proteins with the same domains tend to share similar functions. A common method to identify the function of newly discovered proteins is to transfer annotation from well-characterized homologous proteins that share a common ancestry [401]. Homologous protein identification methods that are purely sequence-based pose difficulties, because of multi-domain architectures and complex evolutionary relationships [466]. Domains are the building blocks for all proteins and present one of the most useful levels at which protein function can be understood [94, 104]. Studies have shown that comparing domain architectures is an effective method for detecting evolutionary relationships [151].

However, these methods are challenged by "promiscuous" domains that combine in many ways to conduct auxiliary functions to the primary role of the protein [34,47,283]. Most domain-based measures treat all domains equally, but "promiscuous" domains are not directly related by homology and should be given less importance. Lee et al [284] developed a Weighted Domain Architecture Comparison (WDAC) of domain architecture similarity that accurately distinguishes between "promiscuous" and conventional domains and improves homology identification. The Inverse Abundance Frequency (IAF) measures the abundance of domains and is derived from the Inverse Document Frequency statistic used commonly in information retrieval [541]:

$$IAF(d) = log_2 \frac{p_t}{p_d} \tag{4.20}$$

where p_t is the number of total proteins and p_d is the number of proteins containing domain d. Highly conserved protein domains appear in a relatively small number of domain architectures, and their neighbor proteins are also conserved during evolution [35]. Domains with identical abundance levels can have different levels of distinct domain partners. The Inverse Versatility (IV) of a domain, d, is defined as:

$$IV(d) = \frac{1}{f_d} \tag{4.21}$$

where f_d is the number of distinct domain families adjacent to domain d. The weight of a domain is calculated by:

$$W(d) = IAF(d) \times IV(d) \tag{4.22}$$

The domain architecture for each protein is constructed using the Vector Space Model (VSM) [429]:

$$x_i = \begin{cases} IAF(d_i) \times IV(d_i) & \text{if } d_i \in p \\ 0 & \text{otherwise} \end{cases} \tag{4.23}$$

where p is the given protein, d_i is the given domain, $d_i \in p$ is whether the given domain exists in the protein's domain architecture, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and n is the number of protein domains in the dataset. The similarity between two protein domain architectures is calculated and returns a range of $[0, 1]$ using the cosine similarity score:

$$sim(X, Y) = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2} \times \sqrt{\sum_{k=1}^{n} y_k^2}} \tag{4.24}$$

where **x** and **y** are vectors of two domain architectures X and Y.

All Homo Sapien genes with annotated protein domains were retrieved from the Interpro [225] database. The WDAC score is calculated for all pairs of annotated Homo Sapien genes. This method was chosen as a more sophisticated approach than a standard cosine similarity score between frequency vectors. In comparison, Lee et al. [284] found that the true positive rate (TPR) for WDAC results (91%) were 6% higher than unweighted domain architecture comparison (85%).

4.4.1.4 Pathways

Disease genes are more likely to share common pathways, because conditions can result from pathway breakdown [250]. Biologically, if any of the pathway components were disabled, they could potentially result in similar phenotypes [58]. Pathway methods are used by various gene prioritization methods [4, 85, 88, 157, 174, 177, 271, 296, 451]. In this study, pathways are retrieved from NCBI BioSystems [175] which contains records from several pathways databases including KEGG [248], BioCyc [64], Reactome [111], Pathway Interaction Database [443], and WikiPathways [257, 388]. After download of the NCBI BioSystems database, an interaction network was created with links for all pairs of Homo Sapien genes with a shared pathway. 12 high count generic pathways were removed, because the records were closer to categories than true biological pathways. These included Signal Transduction, Metabolism, Gene Expression, Immune System, Cell Cycle, Disease, and Developmental Biology.

4.4.1.5 Expression

Gene expression is one of the least biased types of evidence. Information is provided by high-throughput analysis and the expression level of thousands of genes is monitored simultaneously. Gene expression can give valuable information about the when and where that molecular functions are exerted, and the functional relationships between genes can be inferred [132, 402]. Disease genes are expected to be co-expressed, and utilizing expression data in combination of other data sources has been found to increase performance [307]. Gene expression data has been used previously in several gene-prioritization studies [3, 4, 85, 174, 271, 287, 299, 342, 451].

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [88, 156, 235, 464, 483, 514–516] is a biological database of known and predicted PPIs for a large number of organisms that includes direct (physical) and indirect (functional) associations. The latest version 9.1 contains information of roughly 5.2 million proteins from 1,133 species. It imports knowledge of protein associations from databases of physical interactions and curated biological pathways. Protein-protein associations are also derived from functional genomics data using microarray analyses. Across diverse experimental conditions, co-regulation of genes can be an accurate predictor of functional associations [477]. STRING imports functional associations from gene expression data from the ArrayProspector server [236]. Predictions are derived from a variety evidence types including high-throughput experiments, co-expression, literature, genomic context, and PPI databases. A confidence score between 0 and 1 is assigned for each predicted association, and the scores are derived by benchmarking against the common reference set, Kyoto Encyclopedia of Genes and Genomes (KEGG) [249].

In this study, the co-expression network for Homo Sapien genes is developed from the STRING database[1]. A tab-delimited file[2] of all protein links and confidence scores is downloaded from the STRING server. A coexpression network is then built between all Homo Sapien genes where the edges are weighted by the confidence score between 0 and 1.

4.4.1.6 Literature

Text-mining was one of the first approaches for disease-gene prediction, and Pubmed currently has over 24 million records [389]. A computational approach, such as natural language processing [539], has the ability to analyze millions of biomedical articles exponentially faster than any scientist could conduct manually. Biomedical literature is one of the best repositories of information about disease genes. However, gene-prioritization methods can be challenged when processing the texts of biomedical articles, due to lack of a consistent representation of key concepts and biases towards well-studies genes. Several previous approaches have integrated text-mining tools to derive gene-disease associations [4,85,174,177,271,279,287,296,299,342,381,451].

In this study, the literature association network is also created from the STRING database. STRING computationally predicts interactions from genomic features, gene orthology, and text mining of scientific texts. A large body of literature is parsed to search for statistically relevant co-occurrences of gene names and extract semantically specified interactions using Natural Language Processing [437]. A literature association network is built from the text mining confidence scores from the STRING database between all pairs of Homo Sapien genes. The edges in the network are weighted from the confidence score on a scale from 0 to 1.

[1] http://string-db.org/

[2] http://string-db.org/newstring_download/protein.links.detailed.v9.1.txt.gz

4.4.2 Data Integration

An integrated data warehouse developed by BioSPIDA [202] and a Neo4j Graph database are utilized to create six separate adjacency networks on protein interaction, protein domains, pathways, expression, literature, and gene ontology. With a converted relational database of Entrez Gene, 305,597 protein interactions were extracted and associated with 18,089 genes. These interactions came from three source databases: The Human Protein Reference Database (HPRD) [397], the Biological General Repository for Interaction Datasets (BioGRID) [471], and Biomolecular Interaction Network Database (BIND) [27]. HPRD contains human protein-protein interactions, and BioGRID contains protein and genetic interactions. BIND covers high-throughput and manually curated molecular interactions. The integration of these three protein networks created an adjacency matrix with 129,536 linkages between 17,487 genes. The pathway dataset was built by extracting 3,348 pathways from 10,288 genes in Entrez Gene. In the pathway adjacency matrix, an entry was set to 1 if the pair of genes shared a pathway and 0 otherwise. A total of 2,755,903 edges were set in the matrix. Many more edges were able to be generated, due to the usage of the BioSPIDA data warehouse. BioSPIDA integrates BioSystems XML data into relational tables, and genes could be associated to many more metabolic pathways than identified in Entrez Gene. In Entrez Gene, an average of 3.81 pathway annotations were linked to each gene. By integrating NCBI Biosystems, 10,288 genes contained a total of 134,333 pathway annotations with an average of 13.05 annotations per gene. For the ontology network, 197,010 gene ontology annotations were extracted for 14,885 genes. The adjacency matrix created for the ontology network was fully connected, since semantic similarity was calculated between every pair of genes. As opposed to the protein interaction matrix and pathway matrix, entries were not restricted to boolean 0 or 1 values. They ranged from 0 to 1 where 1 is completely identical and 0 is no similarity. The HRSS semantic similarity measure by Wu et al. [530] was used to calculate the score between each pair of genes. HRSS requires graph computations and common ancestry queries that are either not possible or with unrealistic runtimes for a relational database. Neo4j is a highly scalable Graph Database (GD) that natively stores nodes, and it contains an embedded, disk-based, fully transactional Java persistence engine[3] that delivers excellent performance benchmarks [31,125]. To build a Neo4j Graph Database, tab delimited files of all GO concepts and relationships were exported from the BioSPIDA GO relational tables. The tab delimited files were then directly loaded into Neo4j community edition 2.0.1 using their batch inserter[4]. 38,618 nodes and 77,323 relationships were inserted into the Graph Database in 1 second and totaling 24.07 Mb.

Two steps were required to build the protein domain network. First, 20,253 genes were matched to their associated Uniprot [28] identifiers. Then, the Uniprot identifiers are used to link to the Interpro [225] database. Interpro is a database of families and predicted domains of protein sequences. All individual matched signatures from a number of member databases are combined into one searchable resource. Examples include Pfam [36], PROSITE [224], SUPERFAMILY [525], SMART [291], and PANTHER [490]. Although it is advantageous that Interpro is comprised of many signatures from disparate sources, in many cases the predicted domains can overlap. In extracting the domains for each gene, overlapping regions were resolved by removing any smaller entries than were inside others. A total of 68,908 domains were associated to 14,585 genes. The domain adjacency matrix was also a fully connected network, because the domain similarity had to be calculated for every pair of genes. Domain similarity was determined by the WDAC comparison metric developed by Lee et al [284]. The final score in each entry of the matrix ranged from 0 to 1 where 1 was identical and 0 was no similarity.

[3] http://www.neo4j.org/

[4] http://docs.neo4j.org/chunked/milestone/batchinsert.html

The final two matrices used in this study were for gene expression and co-referenced literature. Both of these matrices were built using the STRING [156, 235, 464, 483, 514–516] database. The BioSPIDA data warehouse contains the correlation scores for all proteins in the STRING database. These scores can be linked to Entrez Gene records, since it has ENSEMBL protein identifiers. Each correlation score was also on a scale from 0 to 1. The total amount of nodes (genes) and edges (relationships) for all six matrices can be viewed in Table 4.1. There are slightly fewer genes in each matrix, because any entries without 'live' status in Entrez Gene were excluded.

Table 4.1: Matrix size for each data source in gene prioritization.

Data source	*Genes*	*Annotations*	*Complete graph*	*Representation*
Domains	20,253	2,268,886	205,081,878	1.1%
Expression	14,405	717,215	103,744,810	0.7%
Gene Ontology	16,413	129,379,289	134,685,078	96%
Literature	17971	1,463,983	161,469,435	0.9%
Pathways	9,953	2,755,903	49,526,128	5.6%
Protein interaction	17,487	132,300	152,888,841	0.1%

4.4.3 Parameter Tuning

Before prioritization of candidate disease genes, all matrices for each data source must be converted into diffusion kernels. As described in Section 4.3, diffusion kernels are created by computing the negative graph Laplacian of the adjacency matrix and then taking the matrix exponential. However, the domain, ontology, expression, and literature matrices are all weighted adjacency matrices. Their entries can contain any value between 0 and 1. Therefore, the *weighted graph Laplacian* must be calculated:

$$L_{ij} = \begin{cases} -w_l & \text{if } i \sim j \\ \sum_{l \sim (i,k)} w_l & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{4.25}$$

where $l \sim (i, j)$ means edge l connects nodes i and j, and w_l is the weighted value for edge l.

The matrix exponential $e^{-\beta L}$ contains an input parameter β to control the extent of diffusion. As stated in equation 4.11, if the eigenvalues are computed for the Laplacian, this parameter can be tuned without repeating the expensive matrix exponential operation:

$$e^{-\beta L} = \sum_{i=1}^{n} v_i e^{\beta \lambda_i} v_i^T \tag{4.26}$$

Without computing a matrix exponential on each iteration, cross-validation is used to tune each parameter for all six data type matrices.

Köhler et al. [271] defined 110 disease-gene families[5] on the basis of entries from the Online Mendelian Inheritance in Man (OMIM) [206] database. Using gene-centric assessment, known disease (seed) genes from these families are tested relative to a background gene set. The background gene set is created from the 100 nearest genes to a disease gene according to their genomic distance on the same chromosome. Cross-validation is used to the measure the performance of the training procedure and tune the parameters for each diffusion kernel matrix. The fold enrichment score [271, 299] is defined as the average rank before prioritization divided by the rank after prioritization. For 100 genes, the average rank is 50 and the enrichment score is (average rank)/(prioritization score rank). If the test disease gene is ranked 1, its enrichment score is 50/1 = 50. In leave-one-out cross validation, a test gene is removed from its disease set and placed with its 100 nearest neighboring genes. The prioritization

[5] http://compbio.charite.de/genewanderer/diseaseGeneFamilies.html

score is then calculated for the interval of 100 genes based on their connectivity to the remaining disease seed genes. With the use of the diffusion kernel $\mathbf{K}$, each candidate gene j was assigned according to the following score:

$$score(j) = \sum_{i \in disease\ gene\ family} \mathbf{K}_{ij} \tag{4.27}$$

Leave-one-out cross-validation is repeated for each gene in every disease set and a final assessment score is established. A grid search is performed on parameter values from 10^{-6} to 10^{2}. The final parameter values and highest average enrichment score for each matrix can be viewed in Table 4.2.

Table 4.2: Highest performing parameter for each diffusion kernel.

Data source	β	*Enrichment*
Domains	2.0×10^{-1}	9.39
Expression	6.0×10^{-2}	16.34
Gene Ontology	9.0×10^{-4}	3.87
Literature	3.0×10^{-3}	39.89
Pathways	9.0×10^{-4}	20.92
Protein interaction	8.0×10^{-2}	18.60

4.4.4 Performance Assessment

After the creation of a diffusion kernel for each data type and parameter tuning, prediction scores for candidate genes must be evaluated for each of the 110 disease gene families. There are a total of six data matrices based on ontologies, expression, literature, protein interaction, protein domains, and biological pathways. Also using leave-one-out cross-validation, a receiver-operating characteristic (ROC) curve is developed to assess how well known disease (seed) genes are ranked relative to non-seed genes. The ROC curve plots true positive rate (sensitivity) versus false positive rate

(1 - specificity):

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{4.28}$$

$$1 - Specificity = \frac{FP}{(FP + TN)} \tag{4.29}$$

$$\tag{4.30}$$

Sensitivity is defined as the frequency that a “target” gene is assigned above a given threshold, and specificity is the percentage of genes ranked below the threshold. Therefore, a sensitivity/specificity value of 80/90 would indicate that the correct disease gene is ranked amongst the best-scoring 10% of genes in 80% of the prioritizations [4]. The sensitivity and specificity values are plotted on ROC curves for each data type and the area under the curve (AUC) is used as a standard measure for performance. During initial tests, the results for gene ontology and domains were significantly lower than the other four datasets ($AUC_{go} = 0.53, AUC_{domain} = 0.58$). In comparison, an AUC score from random guessing would roughly equal 0.5. After further investigation, it was found that many of the disease networks in the GO, domain and other datasets were not informative. Similar to the method by Chen et al. [88], this study defines the informativeness I_G^l of a network in disease family G and data type l with a set A of known genes as:

$$DKPC^l(i,j) = \frac{\mid \{(s,t) \mid \mathbf{K}^l(s,t) < \mathbf{K}^l(i,j)\} \mid}{\mid \{(s,t) \mid \mathbf{K}^l(s,t) > 0\} \mid} \tag{4.31}$$

$$I_G^l = \frac{\sum_{i \neq j \in A} DKPC^l(i,j)}{C_{|A|}^2} \tag{4.32}$$

As stated previously, the possible data types of the network are protein interactions, GO, protein domains, pathways, expression, and literature. $C_{|A|}^2$ is the binomial coefficient with parameters $\mid A \mid$ and 2. This equates to the average informativeness between all pairs of genes. In equation 4.31, informativeness between a gene pair is defined as the percent of values with a lower diffusion kernel score.

In analyzing the results for AUC for GO, it was found that only 23/110 diseases received a score above 0.7. Interestingly, only 36 diseases had a score above 0.68 for the informativeness of its network. Other networks exhibited this behavior as well where the AUC scores significantly decreased when the informativeness dropped to lower levels. Several different thresholds were chosen for each network, because networks varied on the required informativeness level before returning acceptable AUC scores. GO was prescribed the highest threshold at 0.68. Diseases below this threshold averaged an AUC score of 0.51, but above returned an average of 0.73. Informativeness only reports the disease network strength in comparison to all other gene pairs. Having a high informativeness does not necessarily imply a high similarity. In comparison, GO was a complete network where edge weights were computed for every gene pair with a metric that could return low similarity measures. Literature did not require a threshold, since all diseases had a high network informativeness and returned a high AUC score. All remaining threshold values for each data type and AUC scores for diseases above and below threshold are reported in Table 4.3.

Table 4.3: Average AUC scores for diseases above and below threshold.

Data source	*Informativeness*$_{\uparrow}$	$\overline{AUC}_{\uparrow}$	$\overline{AUC}_{\downarrow}$	*Diseases*$_{\uparrow}$
Domains	0.58	0.75	0.52	43
Expression	0.58	0.85	0.57	65
Gene Ontology	0.68	0.73	0.52	36
Literature	0.0		0.97	110
Pathways	0.56	0.88	0.66	89
Protein interaction	0.63	0.85	0.58	79

$^{\uparrow}$ Above threshold
$^{\downarrow}$ Below threshold

The network informativeness thresholds are then used to integrate scores from all six data types. Without the thresholds, it would not be appropriate to integrate GO or domain network scores, due to their low AUC score. Li and Patra [293] integrate weighted scores from four separate data types into their final prediction score, but no contribution is permitted from any data source with an AUC score below 0.7. In an experimental setting, a very informative network should in theory make accurate predictions when prioritizing candidate disease genes. However, the actual AUC score would be unknown, because it is not possible to determine true or false positives on prioritizations of new genes that have no existing associations. The informativeness of a network of known disease genes would still be able to calculated. Therefore, in this study, a threshold is selected for each of the six data types based on network informativeness. The AUC score is then calculated to represent how well disease genes are discriminated from its 100 nearest neighbors. This reflects a realistic setting where techniques such as linkage analysis or positional cloning can only pinpoint the responsible area for a disease to a rough location on a chromosome.

The next step is to weight the scores from each data type into a final assessment score. DIR [88] also integrates multiple data types, but only the score from the most informative network may contribute to the final score for each disease. This does eliminate poorly scoring networks from significantly reducing the overall score, but it only allows one data source to contribute for each disease. Li and Patra [293] allow all data sources to factor into the final ranking for a candidate gene, but only if the AUC score is above a 0.7 threshold. The final score is calculated by weighting a discounted rating with the AUC score from each data source. WINGSPREAD also utilizes discounted rating, but uses weights from the informativeness of each data type and a threshold for each network. If the informativeness is below a threshold, the data source will not be used to calculate the final ranking. A discounted rating is calculated by:

$$rating_i = rating_{max} - \left\lfloor \frac{(r_i - 1) \cdot rating_{max}}{N} \right\rfloor \tag{4.33}$$

$$dr_i = \frac{rating_i}{log_2(r_i + 1)} \tag{4.34}$$

where N is total amount of candidate genes (100), $rating_{max}$ is the selected maximum rating, and r_i is the returned prioritization ranking. For example, if the selected maximum rating is 5, and the prioritization method ranks a candidate gene 30th out of 100, the rating score would be $5 - \lfloor \frac{30 \times 5}{100} \rfloor = 5 - 1 = 4$. The rating is then reduced by the ranking number to create a discounted rating score dr_i. In this example, $dr_i = \frac{4}{log_2(30+1)} = 0.82$. Discounted rating originated from discounted cumulative gain (DCG) scoring widely used in information retrieval [234, 293]. In DCG, there is a larger difference in scores between top ranked items than lower ranked items.

In WINGSPREAD, the final weighted discounted rating for a candidate gene in a disease set is calculated as:

$$wdr_i = \frac{\sum_{l \in \mathcal{M}} -log(1 - I_G^l) \times dr_i^l}{\sum_{l \in \mathcal{M}} -log(1 - I_G^l)} \tag{4.35}$$

where $\mathcal{M}$ is the set of all data sources, and I_G^l (equation 4.32) is the informativeness for disease G and data source l. The logarithm is applied to informativeness to help contribute to the stability of the score. Networks with higher informative levels will be weighted more highly. Discounted ratings is also used to give greater scores to candidate genes with higher ranks. Network informativeness thresholds, logarithms of informative weights, and discounted ratings are all used to prevent lower scoring networks from degrading performance.

After the weighted discounted ratings are calculated for all 100 candidate genes, the scores are ordered from highest to lowest. For leave-one-out cross-validation, one of the 100 candidate genes is the removed gene from the disease set. The weighted discounted rating of the removed disease gene is retrieved, and its place in the ordered list is reported (1 is the best possible score and 100 is the worst).

The AUC scores are calculated for all diseases and all data sources, and the most informative networks for each disease are determined. The ROC curves are then plotted for each individual data source as well as for the integrated data sources. The final results are compared to estimate how much the performance raised after integrating multiple data sources versus only utilizing individual sources.

4.5 Results

In analyzing the results for each data source, literature gave by far the highest accuracy with an overall AUC score of 0.96. All 110 diseases scored an AUC over 0.70 for the literature dataset. The AUC scores for all data types as well as the number of diseases above an AUC score of 0.70 can be seen in Table 4.4. GO and domains were the lowest scoring data sources with an overall AUC score of 0.53 and 0.58 and with only 23 and 31 diseases above 0.70 AUC. However, if the ROC is only calculated for diseases above a network informativeness threshold, all AUC scores are raised considerably. This is in exception to literature, because all networks already score above the informativeness threshold. Intuitively, this is an appropriate method for an experimental setting, because realistically any network that is not deemed informative should be disregarded from making predictions. A final score is then calculated by integrating each data source and appropriately weighting their contribution according to equation 4.35. As noted previously, a score from a source will receive a weight of 0 if its informativeness is below threshold. The AUC score when integrating all data sources is 0.9543, slightly higher than the AUC score of 0.9521 for literature, the highest scoring data type. The ROC curves for each data source with and without the network informativeness threshold can be seen in Figures 4.1 and 4.2. Figure 4.2 also includes the ROC curve for the integrated data sources.

Table 4.4: AUC scores for all data sources with and without informativeness threshold.

Data source	*AUC*	$Diseases_{\dagger}$	$AUC_{\uparrow}$	$Diseases_{\uparrow}$
Domains	0.58	31	0.71	43
Expression	0.70	61	0.82	65
Gene Ontology	0.53	23	0.73	36
Literature	0.95	110	0.95	110
Pathways	0.83	92	0.86	89
Protein interaction	0.71	79	0.80	79

↑ Above informativeness threshold
† AUC score > 0.70

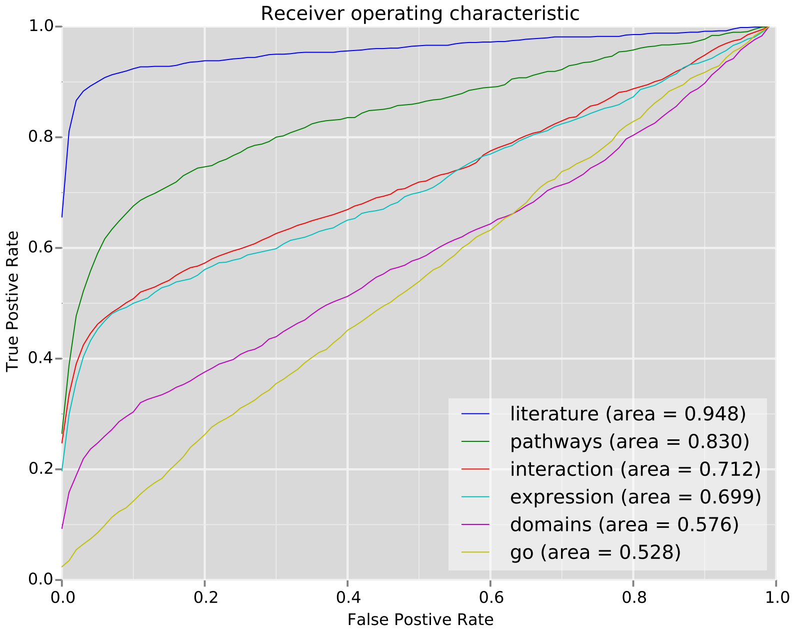


Figure 4.1: ROC performance curves for all six data sources with no network informativeness thresholds.

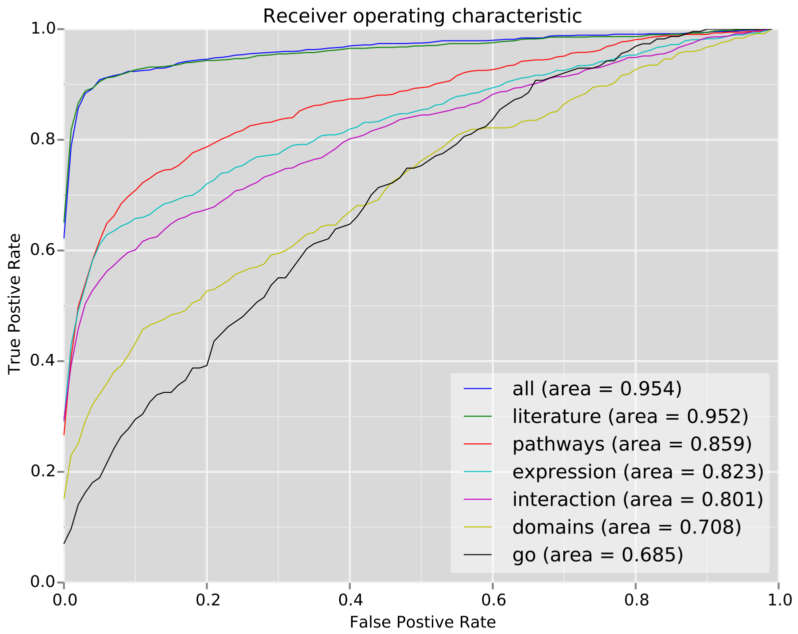


Figure 4.2: ROC performance curves for all six data sources with network informativeness thresholds.

The overall AUC score for integrated data sources is only increased by 0.2% from the score for literature. However, the AUC score for literature is already substantially high without a large area for improvement. Disease genes are identified at top 1/100 ranks in 770/1183 of prioritizations. To compare the integrative strategy, the weighting methods by Li and Patra [293] and DIR [88] are also tested. Li and Patra weight scores from each data source by the AUC score returned for each disease:

$$wdr_i = \frac{\sum_{l \in \mathcal{M}} (\mu_G^l \cdot dr_i^l)}{\sum_{l \in \mathcal{M}} (\mu_G^l)} \tag{4.36}$$

where dr_i is calculated using the same techniques as in equation 4.34 and μ_G^l is the AUC score returned for disease G and data type l.

DIR only uses the score from the data type with the most informative network. This is to prevent from degrading the AUC score from higher performing networks. Intuitively this makes sense, because if literature is known to return high AUC scores, an average with scores from lower scoring networks, such as domain or GO could significantly reduce overall accuracy. The formula for most informative discrimination used in this study is the following:

$$\begin{aligned} l^* &= \arg\max_l \{I_G^l\} \\ wdr_i &= dr_i^{l^*} \end{aligned} \tag{4.37}$$

In testing the integration methods for all six data sources, the informative weighting method in WINGSPREAD (equation 4.35) still received the highest overall AUC score at 0.9543. The AUC weighting method (equation 4.36) received a score of 0.9537. This was in despite of weighting with AUC scores for each data source and disease. As stated previously, in an experimental setting, the true AUC scores for these diseases and data sources technically would not be known in the discovery of new genes. This is because the AUC scores were generated in cross-validation only for known genes, so it is encouraging a higher score was achieved by only weighting the network informativeness of known disease genes. Using the most informative discrimination method (equation 4.37) received a score of 0.9503 which was lower than the highest scoring literature network. This was probably due to the fact that other data source networks could provide a higher informativity score for certain diseases but still resulted in lower prediction performance. The informativeness only depicts how many scores in the network are higher than the similarity values between all known disease genes. It does not give information pertaining to the overall distribution of scores in the network. For example, it was found with the GO network that higher informative scores (percentiles) were needed before reaching acceptable AUC levels. Therefore, most informative discrimination most likely used other networks when they had higher informativeness levels than literature but lower performance. So, the overall AUC score was less than when only using literature.

The most informative discrimination method used by DIR was slightly different than this study. The most informative network is used for each disease gene association rather than for only calculating the overall ranking:

$$score(b) = \sum_{a \in A} \max_{l \in \mathcal{M}} \{DKPC^{l}(b,a)\} \tag{4.38}$$

where A is the set of disease genes, b is the tested candidate gene, and DKPC is derived by equation 4.31.

However, this method also compares network informativeness between data sources when calculating rankings without factoring the actual distributions of the diffusion kernel values. For example, the difference in kernel values at the 80th and 90th percentile could be substantially larger than the kernel values at the 20th and 30th percentile. So, it may not be reliable to only use percentile scores. DIR did not use literature as one of their data sources. In excluding literature, the AUC score using its most informative discrimination method (equation 4.38) was 0.8508. When using the most informative method by this study (equation 4.37), the AUC score was 0.8505. The choice of most informative method did not substantially alter results.

Gene prioritization tools can give optimistic estimates of their predictive power because of "knowledge contamination," since knowledge of disease-gene associations are rapidly integrated into biomedical articles, ontologies, and pathways after publication [226, 343]. In an effort to duplicate a real world experimental setting, the integrative strategy is tested by removing data sources that could introduce biases towards well-studied genes. Pathways, literature, and ontologies were removed from the integrated data sources. With only utilizing interaction, domains and expression, the AUC score using informative weighting was 0.856. This was higher than all individual networks: interaction (0.823), domains (0.710), and expression (0.833). The ROC curves for the integrated data and 3 individual data sources are plotted in Figure 4.3. The AUC scores for individual networks are slightly different than Figure 4.2, because the nearest 100 genes are changed based on available annotations. If a nearest gene does not contain available annotations for all integrated data types, it could potentially inflate the scores by preferencing genes with available annotations. The AUC scores for AUC weighting and most informative discrimination were also both lower at 0.856 and 0.844. This process was repeated by adding other data source combinations including gene ontologies, pathways, or both. Each case returned a higher overall score for informative weighting than the other integrative methods and also a higher score than any individual network. For each case, Gene Ontology did slightly decrease the overall AUC score when added to available data sources. This may be due to generally the lower performance for Gene Ontology. The AUC scores for each weighting method and individual data sources can be viewed for all integration cases in Table 4.5. The ROC curves for the remaining cases of integrated data sources can be viewed in Figures 4.4 - 4.6.

Table 4.5: AUC scores for all combinations of data sources with different integrative strategies.

Data sources	*Informative Weighting*	*AUC Weighting*	*Most Informative Discrimination*	*Diseases*	*Interaction*	*Domains*	*Expression*	*Pathways*	*GO*	*Literature*
INT, DOM, EXP	0.857	0.856	0.844	96	0.823 (79)	0.710 (43)	0.833 (65)			
INT, DOM, EXP, GO	0.854	0.851	0.839	99	0.822 (79)	0.708 (43)	0.832 (65)		0.695 (36)	
INT, DOM, EXP, PATH	0.867	0.859	0.851	105	0.802 (79)	0.701 (43)	0.824 (65)	0.859 (89)		
INT, DOM, EXP, PATH, GO	0.866	0.859	0.850	105	0.801 (79)	0.708 (43)	0.824 (65)	0.859 (89)	0.686 (36)	
INT, DOM, EXP, PATH, GO, LIT	0.954	0.954	0.950	110	0.801 (79)	0.708 (43)	0.823 (65)	0.859 (89)	0.685 (36)	0.952 (110)

All decimal values are AUC scores
Values in parentheses represent the number of scored diseases
INT protein interaction; DOM protein domains; EXP gene expression; PATH biological pathways; GO gene ontology; LIT biomedical literature

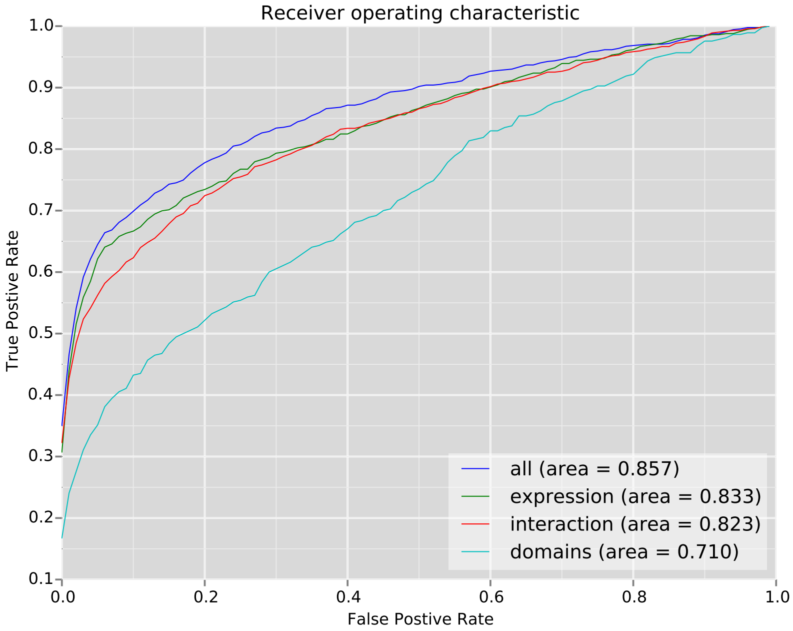


Figure 4.3: ROC performance curves with network informativeness thresholds for protein domains, gene expression, and protein interactions for testing accuracy without "knowledge contamination."

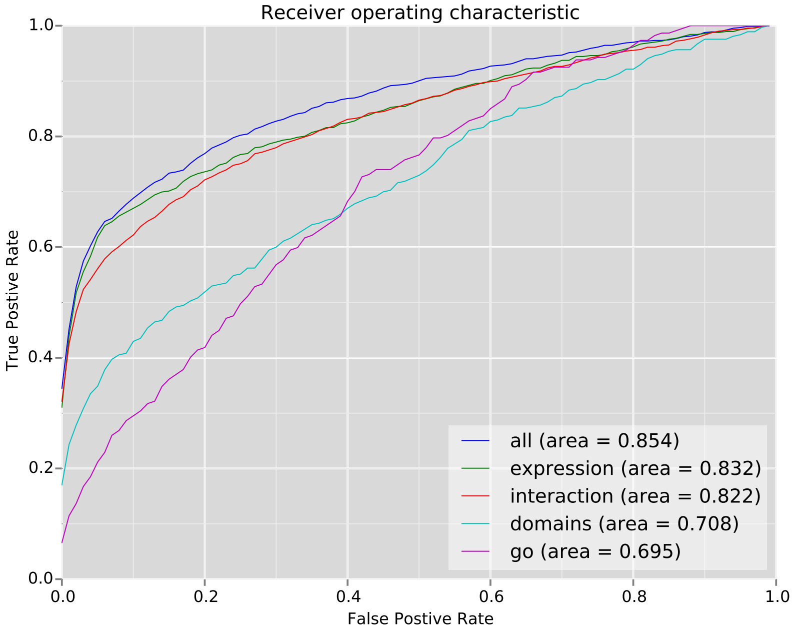


Figure 4.4: ROC performance curves for protein domains, gene expression, protein interactions, and gene ontology with network informativeness thresholds.

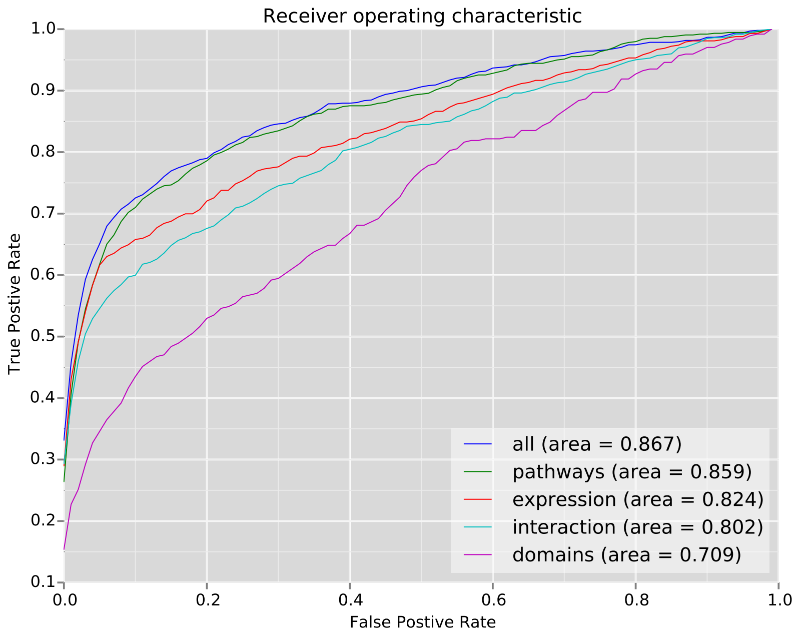


Figure 4.5: ROC performance curves for protein domains, gene expression, protein interactions, and biological pathways with network informativeness thresholds.

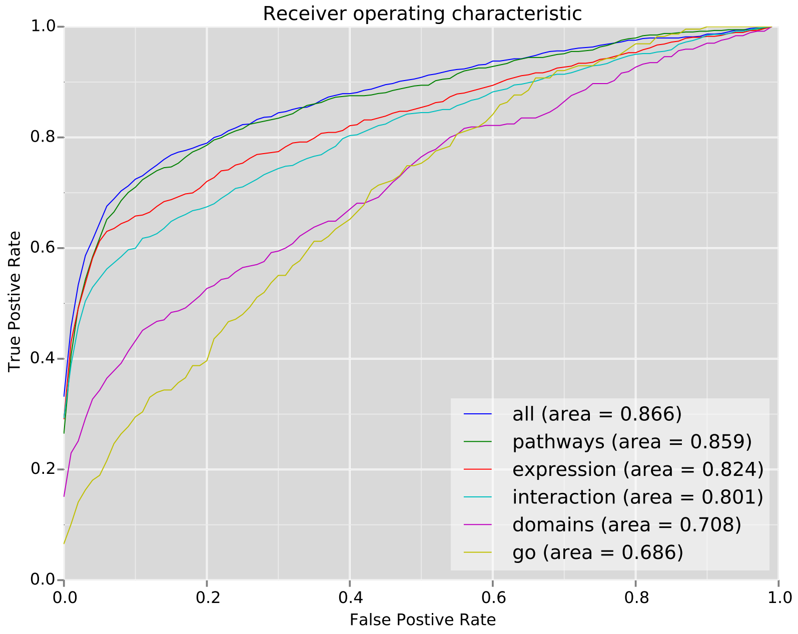


Figure 4.6: ROC performance curves for protein domains, gene expression, protein interactions, biological pathways, and gene ontology with network informativeness thresholds.

In comparison to previous prioritization methods, the overall performance for WINGSPREAD was higher when calculating the AUC score for all 110 disease prioritizations. In the study by Chen et al. [88], the AUC score achieved for the 110 disease families defined by Köhler et al. [271] was 0.80 for DIR. ENDEAVOUR and GeneWanderer were tested as well returning scores of 0.785 and 0.749. WINGSPREAD returns an AUC score of 0.954 when incorporating all six evidence types. Removing the data sources that can be susceptible to knowledge contamination reduces the score, but still results in higher performance than the other competing methods. With only domains, interaction, and expression data sources, WINGSPREAD returns an AUC score of 0.857. This score is only for 96 diseases. The exclusion of diseases is either due to not containing enough annotations for disease genes or network informativeness is not high enough for any of the included data types. As more data types are added, AUC scores are returned for a greater amount of diseases (Table 4.5).

In calculating the AUC for all diseases, the GO and domain data sources did score well below acceptable levels for a predictive model. However, it is illustrative how the diseases with an informative network gave very high AUC scores even for the lower performing data sources of GO and domains. The top 5 scoring diseases for each data source can be observed in Table 4.6. Network informativeness can be viewed as a filtering method, where annotations from certain data sources are very predictive for a particular disease, while they are not as relevant for others.

Table 4.6: Top diseases by AUC score for each evidence type.

Disease	*interaction*	*domains*	*pathways*	*go*	*literature*	expression	all
Interaction							
Cholestasis	**0.99**	0.81	0.73	0.54	0.99	0.66	0.99
Elliptocytosis	**0.99**	0.95	0.90	0.25	0.99	0.99	0.99
Leukoencephalopathy with vanishing white matter	**0.99**	0.22	0.95	0.68	0.99	0.99	0.99
Bare lymphocyte syndrome type II	**0.99**	0.60	0.99	0.82	0.99	0.48	0.99
Adrenoleukodystrophy	**0.99**	0.78	0.99	0.70	0.99	0.82	0.99
Domains							
Hypokalemic periodic paralysis	0.78	**0.99**	0.97	0.50	0.99		0.99
Pituitary dwarfism	0.99	**0.98**	0.56	0.50	0.99	0.85	0.99
Juvenile myelomonocytic leukemia	0.90	**0.96**	0.71	0.33	0.99	0.59	0.99
Elliptocytosis	0.99	**0.95**	0.90	0.25	0.99	0.99	0.99
Noonan Syndrome, Costello syndrome, Cardiofaciocutaneous	0.98	**0.93**	0.94	0.51	0.94	0.69	0.99
Gene Ontology							
Multiple Acyl-CoA Dehydrogenase deficiency	0.78	0.79	**0.99**	0.87	0.99	0.99	0.99
Adrenoleukodystrophy	0.99	0.78	**0.99**	0.70	0.99	0.82	0.99
Achromatopsia	0.64	0.67	**0.99**	0.81	0.99		0.99
Bare lymphocyte syndrome type II	0.99	0.60	**0.99**	0.82	0.99	0.48	0.99
Kartagener syndrome	0.39	0.55	**0.99**	0.58	0.86	0.86	0.91
Domains							
Polycystic kidney disease	0.68	0.91		**0.99**	0.99	0.09	0.99
Pseudohypoaldosteronism, type I, autosomal recessive	0.99	0.67	0.99	**0.99**	0.98	0.99	0.99
Fundus albipunctatus	0.66	0.50	0.98	**0.97**	0.99	0.86	0.99
Cerebrooculofacioskeletal syndrome	0.99	0.74	0.98	**0.92**	0.99	0.92	0.99
Congenital myasthenic syndromes	0.99	0.85	0.82	**0.92**	0.99	0.96	0.99
Literature							
Polycystic kidney disease	0.68	0.91		0.99	**0.99**	0.09	0.99
Fundus albipunctatus	0.66	0.50	0.98	0.97	**0.99**	0.86	0.99
Multiple Acyl-CoA Dehydrogenase deficiency	0.78	0.79	0.99	0.87	**0.99**	0.99	0.99
Hyper-IgM syndrome	0.87	0.47	0.99	0.84	**0.99**	0.63	0.99
Bare lymphocyte syndrome type II	0.99	0.60	0.99	0.82	**0.99**	0.48	0.99
Expression							
Multiple Acyl-CoA Dehydrogenase deficiency	0.78	0.79	0.99	0.87	0.99	**0.99**	0.99
Leukoencephalopathy with vanishing white matter	0.99	0.22	0.95	0.68	0.99	**0.99**	0.99
Maple-syrup urine disease	0.90	0.83	0.99	0.56	0.99	**0.99**	0.99
Elliptocytosis	0.99	0.95	0.90	0.25	0.99	**0.99**	0.99
Stickler syndrome	0.84	0.83	0.98	0.75	0.98	**0.99**	0.99

4.6 Discussion

WINGSPREAD accomplishes gene prioritization by integrating six available data sources on protein interaction, gene ontology, pathways, protein domains, literature, and gene expression. Many previous methods do incorpoate multiple data sources, but usually by focusing on a few databases. It is more common for prioritization strategies to incorporate two to three types of evidence [3, 177, 279, 381, 446, 497, 512, 532, 545], but less frequently are scores integrated from many data sources into an overall ranking [4, 85, 88, 293, 299]. Approaches with integrative methods, such as ENDEAVOUR [4] and ToppGene [85], usually develop a metric for each data source rather than using a unified framework. In a unified framework, new data sources can easily be added without developing a new methodology.

Previous prioritization methods traditionally use metrics to measure associations commonly based on general techniques such as cosine similarity or shared annotations. Li and Patra [293] and BRIDGE [89] use cosine similarity methods for ontology and pathway evidence. Linghu et al. [299] and POCUS [497] use shared annotation scoring for gene ontology and protein domains. As with DIR [88], WINGSPREAD uses a unified framework for integration by representing each data source as a diffusion kernel. Further, sophisticated similarity metrics are used to develop edge weights for each networks. By utilizing the Neo4j Graph database, the HRSS semantic similarity [530] can be quickly computed between GO terms for all gene pairs. This is a more advanced method than vector cosine similarity and has outperformed competing methods in determining functional similarity between gene products. The WDAC metric was used for calculating the protein domain similarity rather than only computing by shared domains between gene products. WDAC was developed by Lee et al. [284], and the study returned better results at identifying homology between multidomain proteins than unweighted domain architecture methods. Using the BioSPIDA [202] data warehouse, multiple protein interaction networks from

BIND, BioGRID and HPRD were able to retrieved and merged into one association matrix. BioSPIDA also provided access to the BioSystems pathway data source. This facilitated significantly greater pathway information per gene than could be obtained by only focusing on the KEGG or Reactome database. BioSPIDA also provided easy linkage to the STRING database to retrieve correlation scores for literature and expression data.

By downloading several databases for each evidence type and utilizing sophisticated similarity metrics, a high AUC score was achieved when integrating all data sources. In all cases of integrated data, the informative weighting method did not lower the overall score. The most informative discrimination method by DIR does prevent from degrading scores from higher performing evidence types. However, it is not possible to achieve a higher overall ranking than returned from the best evidence network. In informative weighting, a high ranking from two evidence types could potentially move the candidate gene to an even higher ranking, since it has multiple support. Informative weighting achieved a higher overall AUC in all data integration cases than most informative discrimination and also AUC weighting. The AUC score for a data source and disease is known beforehand in AUC weighting, and it is used for weighting and as a threshold. However, this would not be realistic in a search for unknown disease genes, and it is encouraging the informative weighting method in WINGSPREAD was still able to return a higher overall AUC score.

The GO and domain evidence types did not return high overall AUC scores when compared to other data sources, such as interaction, literature, pathways, and expression. As can be viewed in Table 4.4, GO and domains did not yield very substantial results for several diseases. GO only received an AUC score above 0.7 for 36 diseases and for 43 diseases for protein domains. This may be due to diseases below this threshold not containing closely related genes in GO and protein domain annotations. These evidence types may also require further testing with other similarity strategies. There is a large variety of semantic similarity measures [107], and their success can depend on the quality of the hierarchical structure of an ontology as well as how balanced annotations are from a given corpus [169]. For example, in GO, many annotations are inferred electronically from sequence similarity of gene products. It is debatable whether these annotations should be included. Many state of the art approaches have proven successful in semantic similarity computations in the biomedical domain [241,378,432,446,505,540], and the best strategy may depend on the given dataset for a particular study.

Other similarity methods are available for protein domain comparison that can incorporate pair-wise comparison of domain architectures [481] or feature architecture similarities [268]. Additionally, methods have been developed to predict domain-domain interactions from sequence [77,409] and structure based models [187,543]. Domain interaction databases also exist, such as DOMINE [405] and INTERDOM [356], that contain known and predicted interactions compiled from a variety of sources.

For gene expression, suspected genes in disease-associated pathways may only be activated in specific cellular compartments or tissues [476]. Tissue specificity may be important in selecting the appropriate gene expression studies when calculating correlation values for a specific disease network. Many complex diseases are associated with changes in gene expression in affected tissues, and certain proteins only interact in specific compartments [239]. BioSPIDA has also integrated access to the Gene Expression Atlas [251,252]. The Expression Atlas provides gene expression patterns under many biological conditions, such as developmental state, organism part, environment factor, and cell type. By focusing on specific expression studies for each disease, the predicted AUC scores could potentially be raised for the gene expression network.

These similarity metrics and databases can potentially be integrated to add stronger and additional evidence for prediction approaches. Future studies can help develop these possible improvements for gene prioritization methods and aid in dissecting the complex genetic architectures behind human diseases.

CHAPTER V

INTEGRATING CLINICAL DATA

5.1 Introduction

Healthcare is a data-rich domain and electronic health records (EHRs) are growing at a rapid pace increasing the demand for efficient data analytics. In 2012, worldwide digitized healthcare data was estimated equal to over 500 petabytes and is forecasted to reach over 25,000 petabytes by 2020 [144]. As the volumes of healthcare data continues to explode, so does its complexity. EHRs are comprised of an enormous variety of data that is structured, unstructured and semi-structured. Unstructured information includes discharge records, doctor notes, and medical imagery such as MRI and CT scans. Structured data can easily be queried and interpreted from machines such as billings and laboratory instrument readings. Despite the steady stream of information from healthcare professionals and medical devices, it is rarely captured and organized into an effective modeling platform for analytics research. Unraveling the complexities of "Big Data" in the medical field can provide many insights to real-time decision making and yield immediate benefits to patients while lowering costs of care [548].

The potential of colossal healthcare data repositories relies on successfully combining primary data with external data sources at both the individual and population level. Often a single data source does not provide the necessary information for accurately conducting a successful analyses for the study in question [103], and it has become essential to supplement existing data by linking to additional data sources. As part of the American Recovery and Reinvestment Act of 2009, the Health Information Technology for Economic and Clinical Health (HITECH) Act provides up to 27 billion in incentives for professionals and hospitals to achieve ‘meaningful use’ of electronic health records. Its goal is to help practitioners and hospitals adopt electronic health records while improving the ability to share and exchange data. Historically, healthcare expenditure in the United States has risen from roughly 9% of the GDP in the 1970s to 17.3% in 2009 [215] without seeing an increase in life expectancy. It is important for the HITECH act to not be another missed opportunity as standard medical practice moves from subjective decision making to evidence-based healthcare.

Without effectively mining and integrating Big Data, the healthcare industry will be in information overload unable to extract new knowledge for providers and patient care. There is a dramatic growth in publicly accessible critical care databases to facilitate informed choices to healthcare practitioners. However, many studies are poorly designed and utilize external datasets that are ill equipped to answer pertinent research questions [489]. To accurately aggregate, manage and analyze big data, healthcare providers must integrate diverse and distributed data while identifying the appropriate data sources for individual studies.

5.2 Related Work

5.2.1 Critical Care Databases

Due to the Meaningful Use legislation from the HITECH Act of 2009, a set of strict terminology standards are defined for use in certified EHRs. These healthcare standards and increased adoption of EHRs have led to a dramatic growth in critical care databases that are easily accessible. Despite the advantages of publicly available health records, there are few references that accurately survey the existing data sources [535].

After researchers have identified the needed variables for an individual study, they must consider several factors when investigating available databases for critical care. These include accessibility, cost, population representation, included variables for each patient, and reliability of coded data. Hospital adoption of electronic medical records has shown promise to improve the quality of intensive care by secondary analyses [103]. The Department of Veterans Affairs created the Inpatient Evaluation Center (IPEC) that includes patient data from over 100 VA hospitals. It includes excellent risk-adjustment measures, but requires partnering with a VA investigator [410]. The Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) [426] database contains 30,000 ICU patients admitted to the Beth Israel Deaconness Medical center. It contains time-stamped treatments as well as minute-to-minute data on physiologic variables. The database contains laboratory results, diagnosis codes, text-based records, discharge summaries, and radiology interpretations. Users can freely gain access to MIMIC-II via the internet[1].

Other types of critical care data sources are created for benchmarking or quality improvement and for administrative purposes. Most benchmarking and quality improvement databases provide clinically rich data to develop risk-adjustment models. Patients voluntarily contribute their data, but the use of the data sources are not universally accessible and usually require a partnership with the provider. The APACHE database contains data from 45 participating hospitals, but access requires a partnership with Cerner [549]. Access to the eICU [330] Research Institute requires partnership with University of Maryland. Administrative data is not as detailed as electronic medical records or benchmarking databases, but they include demographics, admission source, diagnosis codes, length of stay, and billing charges for each hospital stay. The Healthcare Cost and Utilization Project (HCUP) is the largest collection of payer inpatient data in the United States (US) [99]. HCUP's Nationwide Inpatient Sample provides a representative sample of hospitals and patients and can be obtained for a low cost[2]. Other administrative databases from government agencies

[1] http://physionet.org/mimic2/mimic2_access.shtml

[2] http://www.hcup-us.ahrq.gov/nisoverview.jsp

or private groups are more expensive and not as easily accessible [50, 423, 534].

The Informatics for Integrating Biology and the Bedside[3] (i2b2) is an NIH-funded scalable informatics framework that provides clinical investigators a suite of tools for the integration of medical records and clinical research data [350]. i2b2 is built on a Hive framework which is a set of server-side software modules ("cells") that either store data or contain analysis tools [349]. i2b2 provides de-identified EHR data that are downloadable as research datasets[4]. These are a series of discharge summaries from Partners HealthCare listed for natural language processing NLP challenges, including de-identification [502], smoking status [501], obesity [500], medication [503], and relations [504].

5.2.2 Biomedical Terminologies

Many biomedical terminologies were developed as a method to accurately represent knowledge about clinical events and healthcare data. By advancing from clinical notes only composed of free text entries, concepts could be defined without ambiguity and hierarchically classified with subtypes [133]. Clinical vocabularies contain a list of terms that organize and identify healthcare diagnostic and procedural data. In each biomedical terminology specific terms are used to define concepts of its particular domain. There are numerous available medical terminologies with classification systems specific to the application field, such as diagnoses, laboratory measurements, medications, and procedures.

The International Classification of Diseases Ninth Revision Clinical Modification (ICD9-CM) [353] is published by the World Health Organization (WHO) and is the most common classification systems used in the United States for representing patient conditions [134]. It is the accepted standard for classifying morbidity and for creating diagnosis related group (DRG) assignments for billing purposes. The data structure

[3] http://www.i2b2.org/software
[4] https://i2b2.org/NLP/DataSets/

for ICD9-CM is tabular with limited concept hierarchies and no support for semantic linkages. The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) [472] is a more a comprehensive clinical terminology for coding clinical findings, procedures or diseases. It is an ontology that contains concepts, descriptions, and relationships. The 2012AB release totaled 296,237 terms with 539,711 relationships[5]. SNOMED CT is the largest logical terminology in healthcare and is becoming more widely adopted in patient records, but it has its disadvantages [352, 365]. Its logic-based ontology has undergone many design reincarnations and it is more structured to describe patient conditions rather than patient care [205]. The International Shortlist for Hospital Morbidity Tabulation (ISHMT) format is a more generalized compilation of diagnoses created by the Hospital Data Project (HDP) of European Union Health Monitoring Program. It contains a higher aggregation level than ICD-9 with a total of 130 separate diagnoses divided into 21 groups. The official conversion table from ICD codes to ISHMT identifiers is available online[6] and the format has been adopted by adopted in 2005 by Eurostat, the OECD (Organisation for Economic Co-operation and Development) and the WHO-FIC (Family of International Classifications) Network. The Current Procedural Terminology (CPT) [25] is a listing of terms and codes for reporting diagnostic and therapeutic procedures. It is one of the most widely used systems for conducting research, developing guidelines, and for reporting medical services to health insurance carriers [449]. Since, not all supplies, procedures, and services could be coded using the CPT system, the Centers for Medicare and Medicaid Services (CMS) introduced the Healthcare Common Procedure Coding System (HCPCS) [65]. It extends CPT by including codes from external health services such as ambulatory services and durable medical equipment. In HCPCS, all Level I codes are from the CPT terminology set [462]. Level II codes

[5] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT/sourcemetastats.html

[6] http://apps.who.int/classifications/apps/icd/implementation/hospitaldischarge.htm

contain ancillary healthcare services and dental procedures adopted from the Current Dental Terminology (CDT) [22]. The International Classification of Diseases Tenth Revision Clinical Modification (ICD-10-CM) [66] was designed by WHO to provide several enhancements to improve coding accuracy, such as diagnosing multiple conditions, and expanded codes to capture more detail. Unlike ICD-9-CM, ICD-10-CM does not contain a procedure code set [449]. CMS contracted with 3M Health Information Systems to develop the International Classification of Diseases, Tenth Revision, Procedure Coding Set (ICD-10-PCS) [26, 67] to report procedures on inhospital claims. HCPCS and CPT will continue to report procedures on other types of claims after adoption of ICD-10-CM/PCS [259]. The United States Healthcare system will require use of ICD-10 by October 1, 2015 [149]. Logical observation identifiers names and codes (LOINC) [327] represent laboratory test names including chemistry, hematology, microbiology and toxicology. RxNorm [303] is a standardized drug nomenclature developed jointly by the National Library of Medicine (NLM) and Veterans Health Administration.

The Unified Medical Language System (UMLS) [298] was developed by NLM to help healthcare professionals extract and integrate biomedical information from the diverse array of available sources. Often healthcare systems are required to share information and UMLS facilitates the automatic mapping of clinical concepts from different coding systems [48]. UMLS is comprised of three separate knowledge sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon and Lexical Tools. The Metathesaurus arranges mapping between biomedical concepts and the 2014AA Release contains a total of 2.97 million concepts held in over 148 different sources[7]. The Semantic Network represents knowledge in the biomedical domain by defining concepts and the linkages between them. The SPECIALIST Lexicon contains many vocabularies commonly used in biomedical source with syntactic and morphological

[7] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

information. The Lexical tools aid in natural language processing by creating lexical variants, normalizing words, and word indexes. The UMLS knowledge sources greatly benefit the healthcare community by bringing together many clinical vocabularies to aid in the creation of electronic health records, natural language processing, and information retrieval.

5.2.3 Natural language processing

Information technology has transformed health care by capturing huge amounts of patient related information during points of service including diagnosis, laboratory tests, medication, and radiology imaging. This provides an unprecedented wealth of opportunities to biomedical research where sophisticated analysis tools such as data mining, machine learning, and text mining can develop knowledge discovery in patient stratification, drug interactions, disease comorbidities, and clinical outcomes [237]. Despite the enormous potential of the continual large-scale adoption of electronic EHRs, researchers are still faced with the technical challenges of integrating heterogeneous patient data without agreed standards [276].

Many health-care systems in EHRs maintain structured codified data types such as drug prescriptions, laboratory results, and billing diagnoses [269]. However, unstructured clinical text is the most abundant type because it allows medical professionals to freely express detailed assessments and case nuances when generating discharge summaries, radiology reports, pathology reports, family history, and other narratives [332, 422]. The unstructured nature of clinical text provides extra challenges to EHR integration since it is highly heterogeneous and contains many abbreviations, spelling errors, negations and domain-specific references [334]. To derive structured phenotypes from clinical text, phrases and lexical permutations must be recognized with natural language processing (NLP) [427] and mapped to controlled medical vocabularies. For example, associated terms are linked to concepts in RxNorm [303] for medications, Read vocabulary for symptoms, ICD10 [529] or SNOMED CT [472] for diagnoses, or LOINC [327] for laboratory measurements. Typical tasks involved in natural language processing include boundary detection, tokenization, normalization, part-of-speech tagging, shallow parsing, and entity recognition. Examples of NLP-based systems that recognize clinical phenotypes are cTAKES [438], i2b2 HITEx [542], BioMedICUS[8], MedEx [537], MedLEE [160], MedTagger [302], SecTag [119, 120], SymText [267] and MetaMap [16, 17, 137]. There are several surveys that provide a more complete review of available systems [123, 334, 360, 458, 461, 544].

To further the development of integrating large amounts of clinical data, different techniques have been developed to accurately extract clinically relevant features contained in the EHR. Patrick et al. [375] built an automated system to convert clinical notes into SNOMED CT concept. All SNOMED CT tokens were indexed and negation and qualifying terms are identified. Pakhomov et al. [372] match diagnoses to the most frequent examples to a database of 22 million manually coded entries from the Mayo Clinic clinical notes. A Naive Bayes classifier is used for the least certain codes. Kang et al. developed an ensemble voting strategy incorporating seven publicly available concept annotation systems and achieved an F-score higher than any participant in the 2010 i2b2/VA challenge [504]. Adamusiak et al. [2] directly integrated 89 UMLS terminologies directly into a clinical workflow with a powerful relational database. Henriksson et al. [212] utilizes the MIMIC-II database to extract multiword terms and identify synonym relations of SNOMED CT preferred terms with distributional semantics. A more complete review of concept extraction from clinical documents is provided by Meystre et al. [334].

[8] http://code.google.com/p/biomedicus/

5.2.4 Clinical Data Warehouses

Clinical Data Warehouses (CDW) facilitate a unified view of clinical data by consolidating various data sources including medical imaging, discharge summaries, and laboratory results. However, without linking these resources to public repositories with controlled vocabularies, translational medicine will be unable to utilize the rich information from medical guidelines, scientific literature, and biomedical databases for diagnosis and treatment [309, 364]. The adoption of CDWs in hospitals has continuously risen from 64% in 2008 to 45% in 2013 [216], due to their abilities to address various needs of the hospital including administration, patient monitoring [428], and finding significant associations among phenotypes [487]. However, CDWs have currently been unsuccessful in delivering their promises for improved and personalized healthcare largely from conventional data warehouses inability to meet the needs of diverse hospital environments [139, 228]. Biomedical studies require complex integration of highly heterogeneous clinical and research data, and the traditional infrastructure of CDWs are unable to expedite the potential advancements highlighted by the digitization of health records [457]. Many state of the art translational research platforms are currently able to integrate large sets of clinical information with public omics data to further our understanding of disease [61, 210]. Examples of these systems include BRISK [484], caTRIP [324], cBio [69], G-DOC [311], iCOD [456], iDASH [361], and tranSMART [482], SysBioCube [95], STRIDE [305], BTRIS [98], and DW4TR [222]. As mentioned previously, i2b2 is a translational research platform that contains a suite of software modules called cells that can be extended and incorporated into existing systems. Many CDWs rely on the architecture of i2b2 [350] including tranSMART, ONCO-i2b2 [452], and BioSTOR [70]. These high-throughput technologies can analyze a broad spectrum of biological data in the context of clinical care and deliver exploration and visualization capabilities not provided by a conventional CDW. Despite the advancements translational research platforms bring to personalized medicine, the majority of contemporary systems are limited in interoperability and data exchange. Few platforms utilize standard terminologies and ontologies that facilitate the mapping of local data to controlled vocabularies. Without these local alignments, systems have an abridged ability to cross-reference external databases to retrieve additional information. They cannot conduct semantic links to terminologies and exploit subsumption properties that enable progressive computer reasoning. Several translational medicine solutions follow a semantic web approach. Here, links between documents are replaced by links of the interior data elements with added semantics [43]. Through semantic web technologies, these systems can map private data to public resources with controlled vocabularies and automatically infer domain knowledge through semantic reasoning. Examples include IPTrans [340], Corvus [221,326], ASSIST [5], MATCH [460], SESL [407], Receptor Explorer [91], and TCM [80]. However, many of these systems do not take full advantage of semantic web technologies without using existing controlled vocabularies and mapping between local and external resources. Several surveys provide a more complete coverage of available CDW [61,112] and semantic web solutions [81,309] in translation medicine.

5.3 Methods

In this study, the MIMIC-II database, freely available from PhysioNetWorks after application[9], is downloaded as tab-delimited files and imported into a local Postgres 9.1.5 relational database. After completion, all ICU stays and patients are stored with accompanying diagnoses, laboratory readings, procedures, and administered medications. To provide linkages to an ontology, SNOMED CT International Release January 2014 is downloaded from the U.S. National Library of Medicine[10]. To bring

[9] http://physionet.org/mimic2/

[10] http://download.nlm.nih.gov/umls/kss/IHTSDO20140131/SnomedCT_Release_INT_-20140131.zip

together biomedical vocabularies, the 2014AA UMLS Full Release Files are retrieved from the U.S. National Library of Medicine[11]. These contain every vocabulary in the Metathesaurus including a semantic network and natural language processing tools. The core files of SNOMED CT are in tab delimited format and can easily be inserted into relational tables in the Postgres database using the COPY command[12]. These tables and all other other integrated relational tables used in this study can be viewed in Table 5.1. UMLS provides the ICD-9-CM Diagnostic codes to SNOMED CT Map[13] as well as the ICD-9-CM Procedure codes to SNOMED CT MAP[14]. The SNOMED CT diagnosis mapping and procedure mapping both exists as two files each. Each file contains a one-to-one map and a one-to-many map. In the one-to-one map, a single SNOMED CT concept fully represents an ICD-9-CM code while the one-to-many (partial) maps can be broader or narrower than the ICD-9-CM code. Fortunately, both files are in tab delimited format with identical field descriptions. Therefore, they can freely be imported into an identical relational table in the Postgres database. After all MIMIC-II clinical records and SNOMED CT mapping files both exist in the relational database, a single SQL query can easily return matching SNOMED CT concepts for all assigned ICD-9-CM codes during a patient ICU stay. In the cases of one-to-many mappings where multiple SNOMED CT concepts are returned for a single ICD-9-CM code, the SNOMED CT concept with the maximum core usage is selected. The core usage is stored in the mapping files and reflects the percent of total code usage in patient claims data.

[11] http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

[12] http://www.postgresql.org/docs/9.1/static/sql-copy.html

[13] http://download.nlm.nih.gov/umls/kss/mappings/ICD9CM_TO_SNOMEDCT/ICD9CM_-SNOMEDCT_map_201312.zip

[14] http://download.nlm.nih.gov/umls/kss/mappings/ICD9CMV3_TO_-SNOMEDCT/ICD9CMV3_SNOMEDCT_map_201301.zip

Table 5.1: Tables used for concept mapping in clinical data integration.

Table	*Source*	*Description*
icd9cm_snomed_map	UMLS	Mapping from ICD-9-CM diagnosis and procedure codes to SNOMED CT concepts
icd9-cm_diagnosis	UMLS	Full text description of all ICD-9-CM diagnosis codes
icd9-cm_procedure	UMLS	Full text description of all ICD-9-CM procedure codes
mrconso	UMLS	Mapping for each UMLS concept including SNOMED CT
curr_relationship_f	SNOMED CT	Relationships
curr_concept_f	SNOMED CT	Contains Id and active flag for all concepts
curr_description_f	SNOMED CT	Full text description for all concepts
icustay_detail	MIMIC-II	Contains gender, age, and admission id for each patient ICU stay
icd9	MIMIC-II	Contains associated icd-9 diagnosis codes for all patient ICU stays
procedureevents	MIMIC-II	Contains associated procedures for all patient ICU stays
d_codeditems	MIMIC-II	Contains icd-9 procedure codes

In the associated records for each ICU stay in MIMIC-II, there are many diagnoses and procedures that do not have an equivalent mapping to SNOMED CT in the tables provided by U.S. National Library of Medicine. This would result in a substantial loss of clinical information for patient records that cannot match to associated ontological concepts. Metamap is a concept extraction tool developed by the National Library of Medicine for mapping biomedical text to Concept Unique Identifiers (CUI) in UMLS [17]. It uses a variety of natural language processing techniques including tokenization, boundary detection, part-of-speech tagging, and word-sense disambiguation [18]. The September 2013 release was downloaded[15] and

[15] http://metamap.nlm.nih.gov/MainDownload.shtml

installed locally to perform concept extraction on unlinkable ICD-9 codes. Metamap was executed on the diagnosis or procedure description for each ICD-9 code. However, the majority of words and sentences in descriptions for ICD-9 codes in the MIMIC-II database are heavily abbreviated and truncated. The long descriptions for ICD-9 diagnosis and procedure codes are available in excel files online from Centers for Medicare and Medicaid Services[16]. These excel files can be directly imported as relation tables into the Postgres database. Metamap returns UMLS concepts with associated scores and semantic types after its execution on long descriptions of ICD-9 codes. Potential UMLS concepts are returned by Metamap with accompanying scores and semantic types. The maximum score is 1000.0 and higher scores imply a greater relevance of the UMLS concept to the biomedical text. The 2013 release of MetaMap can return up to 133 distinct semantic types, and it is important to exclude UMLS concepts with semantic types not relevant to diagnosis or procedures. Table 5.2 provides all accepted semantic types during execution of Metamap. The UMLS concept with the maximum score and approved semantic type is selected as the matching entry for each ICD-9 code.

Table 5.2: Accepted semantic types returned by Metamap for linking ICD-9-CM codes to SNOMED CT.

Diagnosis	*Procedure*
Disease or Syndrome	Therapeutic or Preventive Procedure
Sign or Symptom	Diagnostic Procedure
Pathologic Function	
Finding	
Injury or Poisoning	
Congenital Abnormality	

[16] http://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/Downloads/ICD-9-CM-v32-master-descriptions.zip

After all matching UMLS concepts are returned for unlinkable ICD-9 codes, the UMLS concepts must then be mapped to SNOMED CT. Metathesaurus in UMLS provides a concept names and sources table (MRCONSO) in Rich Release Format (RRF) [220] and contains a string or text label (STR) for each SNOMED CT Atomic Unique Identifier (AUI). The fields in this file are pipe delimited and all records can be imported into a relational table in the Postgres database. The CUIs returned from Metamap are sent to an SQL query that retrieves the accompanying records from the MRCONSO table. The MRCONSO table has several flags that helps select the most appropriate SNOMED term for each UMLS concept. TTY is the term type in source. It can have many possible values but only SNOMED CT terms with FN or OAF are permitted for retrieval. FN is defined as the full form of descriptor and OAF is defined as an obsolete active full specified name. Other flags that exist in the MRCONSO table are SUPPRESS and ISPREF. SUPPRESS is a suppressible flag with four possible values, but only entries with Y and N are retrieved. Y is non-obsolete content deemed suppressible and N is defined as not obsolete. ISPREF is the atom status and values of Y or N mark whether the entry is preferable. Combinations of values for these flags create a precedence structure in retrieving mappings from UMLS concepts to SNOMED CT. Entries that are preferable and not suppressible with FN term type are selected first. If these do not exist, entries with FN term type and without the preferable flag are returned. OAF term type entries are selected only if no FN entries can be found. The SNOMED CT Concepts table contains all concepts with identifier numbers and an active flag. After all matching SNOMED CT concepts are returned, they are joined to the SNOMED Concepts table (curr_concept_f in Table 5.1) in the database and all items not marked as active are filtered. This process repeats itself for each patient ICU entry in MIMIC-II and its workflow can be viewed in Figure 5.1.

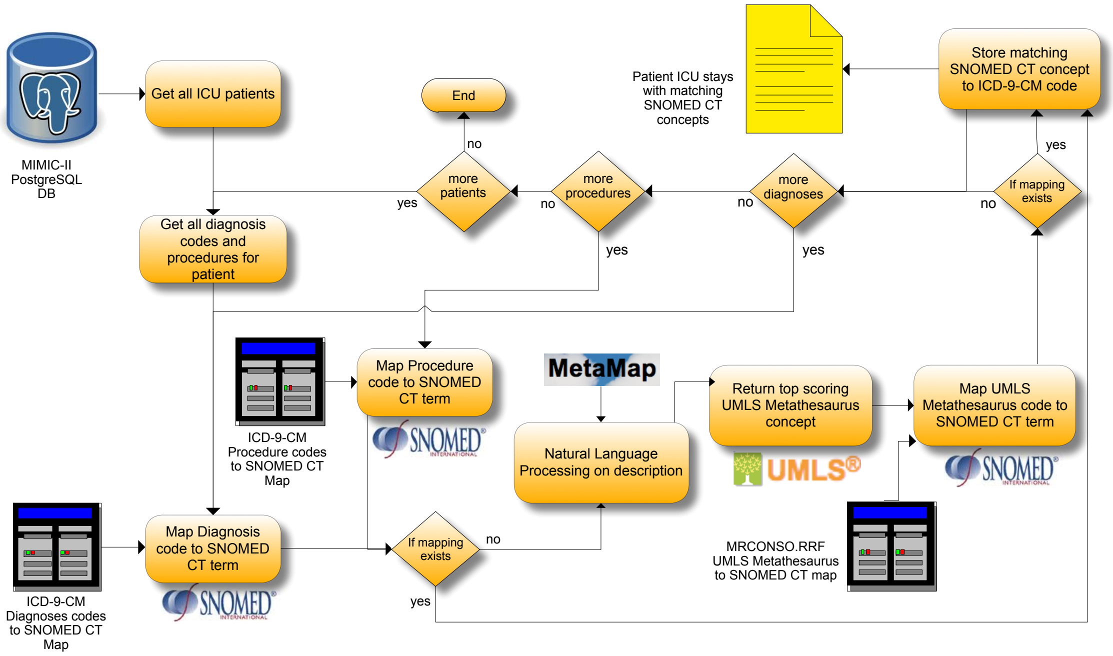

MIMIC-II PostgreSQL DB
Get all ICU patients
End
no
Patient ICU stays with matching SNOMED CT concepts
Store matching SNOMED CT concept to ICD-9-CM code
yes
more patients
yes
no
more procedures
no
more diagnoses
no
If mapping exists
Get all diagnosis codes and procedures for patient
yes
yes
ICD-9-CM Procedure codes to SNOMED CT Map
Map Procedure code to SNOMED CT term
SNOMED
MetaMap
Natural Language Processing on description
Return top scoring UMLS Metathesaurus concept
UMLS®
Map UMLS Metathesaurus code to SNOMED CT term
SNOMED
ICD-9-CM Diagnoses codes to SNOMED CT Map
Map Diagnosis code to SNOMED CT term
SNOMED
If mapping exists
no
yes
MRCONSO.RRF UMLS Metathesaurus to SNOMED CT map

Figure 5.1: Process for linkage of diagnoses and procedures from MIMIC-II clinical entries to SNOMED-CT concepts.

By integrating MIMIC-II, SNOMED CT, and UMLS mappings into a local relational database, biomedical terminologies can quickly be mapped to associated concepts in the SNOMED CT ontology with powerfully customizable SQL queries. However, relational databases have many limitations in the biomedical domain because they deliver unimpressive performance for applications that deeply analyze relationships between entities [125]. They are increasingly disadvantaged when required to traverse many nodes of a ontological graph. Relational databases use foreign keys to perform table joins with logical pointers. They must conduct an index lookup for each edge traversal and this is more expensive depending on the size of the dataset. Graph Databases have maintained to continually attract attention due to a more natural modeling of data and a larger emphasis on interconnections between entities and relationship properties [9, 418]. They are able to directly traverse to neighboring nodes by dereferencing physical pointers negating the need for index lookups. For certain connected data operations, they are reported to reach speeds with execution times one thousand times faster than relational databases [7, 38].

To accurately compare concepts in SNOMED CT, it was essential to additionally store the ontology in a Graph Database (GD). Neo4j is a highly scalable GD that natively stores nodes, relationships and attributes in a property graph model rather than relying on a relational layout of the data. It is quickly becoming one of the most popular choices due to its free open-source GNU General Public License (GPL), and it contains an embedded, disk-based, fully transactional Java persistence engine[17] that delivers excellent performance benchmarks. [31, 125] To build a Neo4j Graph Database, tab delimited files of all SNOMED concepts and relationships were exported from SNOMED CT Postgres relational tables. The tab delimited files were then directly loaded into Neo4j community edition 2.0.0 using their batch inserter[18]. However it was found that in the creation of a SNOMED Graph Database, there were many cycles in the relationships between nodes. This greatly impedes many graph operations such as returning all paths between nodes. By utilizing the Postgres relational database, a SQL query quickly found cycles including only 2 nodes with a double join of the SNOMED Relationship table. Finding cycles containing more than 2 nodes was unrealistic with a relational database due to manual construction of table joins and expensive index lookups. Therefore, no cycles were removed from the SNOMED relational database, and it was extracted and imported into Neo4j creating a directed ontological graph. With the Graph Database, powerful queries could be rapidly executed using Cypher, the declarative graph query language of Neo4j. These included node operations such as shortest path, depth level, and common ancestor queries which were either not possible or unrealistic to execute on a relational database. The Cypher query language was quickly able identify all cycles with 2 or more nodes. To remove cycles, an edge was selected for deletion based on multiple

[17] http://www.neo4j.org/

[18] http://docs.neo4j.org/chunked/milestone/batchinsert.html

criteria. Neo4j calculated three different features for all nodes involved in cycles: maximum depth, shortest path, and fan-in count. Maximum depth is the maximum amount of nodes between the current node and head node of the ontological graph. Shortest path is the shortest distance (number of edges) that can be traveled from the current node to the head node. Fan-in count is the number of incoming edges to the current node in the cycle. In each cycle, edges were removed if the outgoing node had a smaller maximum depth than the incoming node. This is to prevent edges from connecting a shallow node to a deeper node in the cycle. There are cases where the maximum depth of all nodes in the cycles are equal. Here, edges are removed based on the node's shortest path distance to the head of the ontological graph. If the nodes for all edges in a cycle have equal maximum depth and shortest path distance, edges are selected for removal based on fan-in count. An edge is eliminated if an outgoing node has a larger fan-in count than the incoming node. Nodes with larger degrees of incoming nodes are expected to be more generalized. Therefore, an edge in a cycle is removed if it connects to a node with a smaller fan-in count. Traversing upward in the ontological graph should increase generalization rather than specificity. After all selected edges were eliminated, a new tab delimited file was extracted from the Postgres database of all SNOMED relationships without cycles. The current Neo4j database was emptied and reloaded with a directed acyclic graph of the SNOMED ontology. By utilizing a relational database, graph database, and natural language processing, a powerful clinical system was constructed able to conduct term mapping, rapid node traversals, and concept extraction on patient records.

5.4 Results

The local storage of a relational database containing all ICU records in MIMIC-II, SNOMED concepts, and UMLS mappings required modest storage sizes. The MIMIC-II database totaled 244 million records with 33 gigabytes (GBs) in row size and 26 GB in index size. The SNOMED CT relational database and UMLS mapping table (mrconso) totaled 4.8 GB and 2.5 GB. Storage size statistics for all tables in the Postgres relational database can be viewed in Table 5.3. After all data was loaded into the relational database, an SQL query could be constructed to map patients into ISHMT format diagnoses. The runtime for this query was 23.2 minutes, and patient distributions were generated for ISHMT ICD groups and for higher level ISHMT ICD chapters (Tables 5.4 and 5.5).

Table 5.3: Storage sizes for integrated clinical relational database.

Database	*Index size*	*Row size*	*Total size*	*Tables*
MIMIC-II	33 Gb	26 Gb	60 Gb	40
SNOMED CT	3.2 Gb	1.6 Gb	4.78 Gb	9
UMLS	1.5 Gb	992 Mb	2.51 Gb	4

Table 5.4: ISHMT ICD Group frequencies of MIMIC-II patient ICU stays.

Diagnosis	*Total*	*Frequency*
Hypertensive diseases	14511	0.39
Conduction disorders and cardiac arrhythmias	10186	0.27
Ischaemic heart disease	8788	0.24
Heart failure	8185	0.22
Complications of surgical and medical care	7981	0.21
Anaemias	7784	0.21
Diabetes Mellitus	7637	0.21
Liveborn infants	7372	0.20
Renal failure	7176	0.19
Pneumonia	4500	0.12
Chronic obstructive pulmonary disease and bronchiectasis	4120	0.11
Acute myocardial infarction	3964	0.11
Septicaemia	3637	0.10
Cerebrovascular diseases	3265	0.09
Diseases of oesophagus	3173	0.09
Mood disorders	2012	0.05
Angina pectoris	1875	0.05
Mental and behavioural disorders due to alcohol	1852	0.05
Intracranial injury	1759	0.05
Pulmonary heart disease & diseases of pulmonary circulation	1658	0.04
Asthma	1146	0.03
Dyspepsia and other diseases of stomach and duodenum	1078	0.03
Paralytic ileus and intestinal obstruction without hernia	984	0.03
Atherosclerosis	937	0.03
Intestinal infectious diseases except diarrhea	936	0.03
Diseases of pancreas	805	0.02

Table 5.5: ISHMT ICD Chapter frequencies of MIMIC-II patient ICU stays.

Diagnosis	*Total*	*Frequency*
Diseases of the circulatory system	24267	0.65
Endocrine, nutritional and metabolic diseases	18907	0.51
Factors influencing health status and contact with health services	18861	0.51
Injury, poisoning and certain other consequences of external causes	14701	0.40
Diseases of the respiratory system	14395	0.39
Diseases of the genitourinary system	11015	0.30
Diseases of the digestive system	10860	0.29
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	10029	0.27
Diseases of the blood and bloodforming organs and certain disorders involving the immune mechanism	9792	0.26
Mental and behavioural disorders	7143	0.19
Diseases of the nervous system	5149	0.14
Certain conditions originating in the perinatal period	4959	0.13
Neoplasms	4673	0.13
Certain infectious and parasitic diseases	4583	0.12
Diseases of the musculoskeletal system and connective tissue	4199	0.11
Diseases of the skin and subcutaneous tissue	3317	0.09
Congenital malformations, deformations and chromosomal abnormalities	2169	0.06
Diseases of the eye and adnexa	1442	0.04
Diseases of the ear and mastoid process	219	0.01
Pregnancy, childbirth and the puerperium	196	0.01

Utilizing Postgres and Metamap, many ICD-9 diagnoses and procedures from patient records were mapped to associated SNOMED CT concepts. MIMIC-II version 2.6 contained a total of 37,130 ICU stays with 335,366 diagnoses. These contained 5,645 unique diagnoses where 4,800 could be directly linked using the ICD-9-CM Diagnostic codes to SNOMED CT Map provided by UMLS. The remaining 845 diagnoses were executed by Metamap in 22.5 minutes. Metamap returned 572 UMLS concepts of valid semantic types. These UMLS concepts were queried against the mrconso table in the Postgres relational database and returned 519 mappings to SNOMED terms that were not obsolete or suppressed. In total, 5,319 of 5,645 ICD-9 diagnoses were able to be successfully mapped to SNOMED CT. This process was repeated for patient procedures. 33,709 ICU stays had assigned procedures with 1,750 unique entries. 1,199 were able to be directly linked using the ICD-9-CM Procedure codes to SNOMED CT Map. Metamap was executed on the remaining 572 entries in 4.6 minutes and returned 444 UMLS concepts with valid semantic types. After querying against the mrconso table, the UMLS concepts returned 223 mappings to SNOMED CT that were not obsolete or suppressible. A total of 1,422 out of 1,750 unique procedures were successfully mapped to SNOMED CT. The Graph Database of SNOMED CT was quickly constructed with the Neo4j Batch Inserter. 349,948 Nodes and 740,968 Relationships were loaded in 7 seconds totaling 243.5 Mb in storage. Connected data operations could be quickly performed to identify and remove cycles. A Cypher query assigned the depth level to nodes. The runtimes up to depth level 4 can be seen in Table 5.6. Neo4j was able to assign the depth level to thousands of nodes in under 11 seconds for each operation. The shortest paths for 1,420 nodes involved in cycles was calculated in 19.9 minutes. A total of 756 two node cycles and 115 many (three or more) node cycles were found when analyzing the SNOMED Graph Database. For two node cycles, 400 edges, 149 edges and 169 edges were able to be removed based on maximum depth, shortest path and fan-in count. The edges for the remaining 38 cycles were removed from manual inspection. For many node cycles, 79 edges, 5 edges, and 31 edges were removed from maximum depth, shortest path, and fan-in count. After removal, the Neo4j database could be quickly emptied and reloaded with a SNOMED Graph Database with no cycles.

Table 5.6: Neo4J Cypher query runtimes for assigning node depth levels.

Depth level	*Nodes*	*Runtime*
1	20	2839 ms
2	523	4724 ms
3	11,797	7694 ms
4	30,898	10704 ms

5.5 Discussion

This current system utilizes many state of the art techniques with natural language processing and Graph Databases to develop a data warehouse that can accurately map descriptions to standard biomedical terminologies. With Metamap, concepts can be extracted and recognized into UMLS to enable linkage to many other biomedical vocabularies. The Metathesaurus facilitates the conversion of UMLS identifiers into other ontologies and is not only limited to SNOMED CT. Future implementations can take advantage of other repositories by converting laboratory measurements into LOINC, medications into RxNorm, medical imagery into DICOM, and procedures into CPT. Furthermore, natural language processing can be used to analyze and recognize concepts in the unstructured data more deeply by incorporating the detailed notes in discharge summaries, radiology reports, pathology reports, and other narratives.

CHAPTER VI

PRIORITIZING ICU PATIENTS

6.1 Introduction

The current climate of critical care has a heavy challenge to meet growing patient demands while hospital capacity continues to shrink at an alarming rate. According to American Hospital Association, the number of hospital beds has reduced by almost 25 percent in a period of 20 years [23]. Due to Certificate of Need (CON) regulations, an average occupancy level of 85 percent was required before approval to increase capacity [57]. Since, many hospitals had average occupancy below these rates, there was an impression in the health care community that there was excess capacity. For nonprofit hospitals, average rates had reached as low as 66 percent [191]. Consequently, available beds have continued to decrease across states.

In April 2002, a Lewin Group survey reported 62 percent of U.S. hospitals reached or exceeded maximum operating levels. The percentages raised to 79 percent for urban hospitals and 82 percent for level I trauma centers [24]. The Center for Disease Control reported the number of annual emergency department (ED) visits climbed by almost a quarter for the decade ending in 2002. The number of EDs reduced by 15% for the same period [322].

Setting hospital capacity by focusing on occupancy levels has led to serious circumstances. There have been access blocks and substantial increases in waiting times [329]. The relationship between waiting time and average occupancy is not linear. At a point, the average delay can start to rise exponentially to even small increases in utilization [190]. Wait time is dependent on the time between arrivals and service duration. These measures have significant variability, and delays can be considerably different for identical utilization levels. It is not sufficient to only emphasize average occupancy levels when evaluating the process flow of a health care center.

Increasing average wait times for medical care has led to complications that are more significant than economic incentives. Poor patient flow has been found to be associated with elevated mortality rates, longer length-of-stay, and heightened readmission [72, 412]. Sprivulis et al. linked ED overcrowding to a 30% relative increase in mortality [468]. Chalfin et al. identified delays to intensive care were correlated with longer lengths of stay and higher mortality [71]. During periods of stress, a decision to admit a patient may not be entirely clinically driven and nurses are prone to medical errors [300].

Early discharges are more likely at high occupancy levels. The average length of stay can be reduced up to 16% for patients discharged from a busy intensive care unit (ICU) [256]. However, the likelihood of returning increases substantially [129, 256, 465]. KC et al. found overall bounce-back probability was 14% but rose to 37.4% for early discharged patients [256]. Higher severity patients are associated with longer revisit stays raising their net total length of stay. These factors effectively reduce hospitals peak capacity, because the readmission loads add unexpected flow related stresses [300]. Readmitted patients have also been found to have higher mortality rates in addition to longer length of stays. Snow et al. identified mortality rates for returning patients were 26%, three times the general population for surgical intensive care units [465]. Readmissions from premature discharge can increase costs and lead to overall worsening of medical conditions for patients [24].

It is essential to improve the process flow of health care centers with motivations that are not purely economic. The demand for intensive care is high. Green et al. determined 90% of ICUs in New York have insufficient capacity to provide proper medical care [191]. While economics tend to favor high occupancy [183], the quality of care does not. This paper evaluates different priority methods to minimize waiting times for admission to intensive care units. An emphasis is placed on the severity of medical conditions rather than exclusively focusing on market factors. The goal is to maximize the quantity of patients served without sacrificing the quality of care.

6.2 Related Work

In a perfected system, all patients would arrive at the same rate and all patients would have the same condition that requires identical service time. This system would be 100% efficient as many automated manufacturing plants [300]. This is not the case in the health care community. Patients arrive unexpectedly with an immense diversity of conditions. Therefore, it is necessary to optimally fit the distributions for patient arrival and service time. In most studies, the inter-arrival times are regarded as a negative exponential distribution [413]. The length of stay (LOS) can have different distributions for different patient types [272]. The fit distributions can vary from exponential, negative exponential, log-normal, or Weibull [100, 329, 454, 496]. Kokangul et al. applied a Kolmogrov-Smirnov test on five years of admissions to a teaching hospital and found arrivals distributed as a Poisson process and LOS distributed as log-normal [272].

Siddharthan et al. classified patients into emergency and non-emergency care [459]. After collecting data from an emergency department in Florida, patients were grouped as emergency care for major trauma, critical care, minor trauma, and non-critical care cases. Non-emergency care was classified only for primary care patients. 53.3 percent of patients were found to require emergency care and 46.7 percent were non-emergency. The average arrival rate, service rate and waiting time were calculated for both types. The study assumes arrivals follow a Poisson probability distribution and service rate follows an exponential distribution. Using a proper priority queue discipline, [233] it found the average waiting time to reduce by 10 percent for all patients. The queue gave highest priority to emergency care patients, because they had the larger average service time.

Chan et al. utilized a more sophisticated priority queue with 9 categories of patients [73]. Each category is classified by low, medium, or high LOS and by low, medium, or high severity. All groups of patients are tested with three different priority models. The model assumes a patient must be discharged for new arrivals if intensive care units are at full capacity. This is due to the inherent urgency of intensive care. Each priority model enforces the discharge order for patients in intensive care. The three models are based on lowest nominal length-of-stay, smallest probability of readmission, and lowest readmission load. Readmission load is return probability multiplied by average LOS for successive visits. The study results reported the readmission load model outperformed all other priority schemes by up to 10%.

Dobson et al. attempts to accurately estimate the expected number of patients transferred to accommodate more critical arrivals [124]. The study does not use a complex priority scheme compared to Chan et al. Instead, patients are more simply discharged by lowest remaining length of stay. A Markov model is utilized to study the effects of ICU workload on patient bumping.

The difficulty of assigning priority to ICU admissions is to correctly identify the severity of incoming patients. Escobar et al. assess the severity of each patient by assigning the probability of mortality based on sex, age, primary condition and chronic ailments [136]. 16,090 ICD admission diagnoses were grouped into 44 broad categories. Graham et al. used a simpler approach by classifying a diagnosis into high, medium, or low risk [189].

Adding to the complications of accurately identifying patient severity, clinicians typically write diagnosis records in free-textual format. There have been successful attempts to use machine learning and natural language techniques to correctly associate notes with hierarchical codes, such as SNOMED-CT® and ICD-9 [79, 116, 372, 375, 386, 424]. However, these methods have been found to have considerably lower performance in data poor cases [424]. More successful results were attained when a large volume of clinical reports, laboratory results and follow-up reports were available.

Regarding strategies for analyzing ICU workflow, Chan et al. only prioritized patients by how they were bumped from the ICU rather than admitted. Discharges were enforced by attempting to minimize readmission load according to several factors, including return probability and LOS. Dobson et al. also prioritized patient transfers from the ICU [124]. They were ordered according to their remaining length of stay. Both of these studies used sophisticated priority schemes, but were not entirely realistic. Patients were automatically admitted when requesting ICU entry by bumping lower priority patients. However, in healthcare settings it is not uncommon for average wait times for an ICU to exceed 4 hours [321], and bumping patients can cause significant medical complications [129, 256, 465].

6.3 Methods and Computational Design

6.3.1 Data preparation

32,531 medical records were retrieved from a large urban hospital over a one year period from March 2010 to April 2011. Each record included the patients id, registration number, diagnosis, and entrance and exit times of each reserved room during the entire hospital stay. Five separate intensive care units were analyzed for this study: Cardiovascular (CV) Surgery, Neurosurgery, Medical, Neuroscience, and Surgical. Since the distribution of LOS may vary among different patient types [272], Matlab® 8.1 was used to fit the LOS distribution for each ICU.

Of 5,465 hospital ICU visits, 813 contained a missing entry. 14.8 percent of records included the time a patient exits an ICU room without the time of entry. These offending records were temporarily removed to calculate the LOS distributions for each ICU unit (Table 6.1). The fitted distributions were then used to sample entrance times for the records with missing entries.

Table 6.1: Length of stay distribution.

ICU	*LOS Distribution*
CV Surgery	LogNormal (3.89, 0.94)
Neurosurgery	GeneralizedExtremeValue (0.61, 13.50, 24.21)
Medical	LogLogistic(3.71, 0.42)
Neuroscience	GeneralizedExtremeValue(0.51, 14.8, 22.6)
Surgical	GeneralizedExtremeValue(0.64, 21.0, 31.4)

Arrival rates were calculated after all hospital ICU visits contained complete records for entry and exit. Full lists of entrance times were generated, and distributions were fitted from interarrival times for each ICU. Arrivals were separated by emergency and scheduled surgery admissions (Tables 6.2, 6.3). Other statistics were calculated to help identify the process flow of patients through the system. These included return rates after a patient leaves an ICU and after a patient is forcibly bumped from an ICU. Mortality rates were determined for patients entering an ICU (Table 6.4).

Table 6.2: Emergency arrival distribution.

ICU	*Emergency arrival distribution*
CV Surgery	Exponential (31.11)
Neurosurgery	Gamma (0.82, 39.88)
Medical	Weibull (10.65, 0.88)
Neuroscience	Exponential (72.23)
Surgical	Gamma (0.82, 19.15)

Table 6.3: Scheduled arrival distribution.

ICU	*Scheduled arrival distribution*
CV Surgery	Gamma (0.76, 99.64)
Neurosurgery	Weibull (24.52, 0.71)
Medical	Gamma (0.33, 4,746.4)
Neuroscience	Exponential (72.72)
Surgical	Gamma (0.70, 51.92)

Table 6.4: Probabilities for ICU returns and mortality.

ICU	$P(R)$	$P(M)$	$P(R \mid t)$
CV Surgery	0.06	0.34	0.33
Neurosurgery	0.13	0.32	0.42
Medical	0.03	0.34	0.43
Neuroscience	0.06	0.36	0.27
Surgical	0.09	0.13	0.33

$P(R)$ Return probability from icu exit
$P(M)$ Mortality probaility
$P(R \mid t)$ Return probability after early discharge

6.3.2 Natural language processing of clinical diagnosis records

The medical records obtained were not comprehensive enough to conduct a full contextual analysis. In particular, the diagnoses from patient records received did not contain standardized codes, such as ICD-9. They were free text entries ranging only up to 54 characters at maximum. This limits text analysis for each record to a few words at most, but it is useful to test the applicability of natural language processing when the content is very minimal.

Due to the difficulty of uniquely matching a patient's diagnosis with minimal content and non-restricted entries, the goal is to instead classify the severity of a patients condition based on these free text entries. Severity is calculated by identifying key words shown to have high prevalence in cases of mortality.

Of 2,950 diagnoses, 486 resulted in mortality. The clinical terms used in mortality cases were treated with higher severity. A list of words was generated from all diagnosis records. Another list was produced only from the mortality records. NLTK, a natural language processing toolkit for Python, was used to tokenize the words in each list [44]. It was important to only include words in the English dictionary and remove any common stop words. Wordlist is a corpus included in NLTK that contains 234,943 unique English words, and the English Stopwords corpus contains 127 unique words. These corpora facilitate more significant words to be identified in diagnosis records, but many medical terms may be improperly excluded. It is possible that common words used by clinicians are not included in the standard English dictionary provided by the NLTK library.

SNOMED-CT® is a standardized reference that contains millions of medical concepts developed by the American Pathologists and the United Kingdoms National Health Service [472]. The July 2011 release contained 988,921 unique medical terms. We use this release to augment the list of English words provided by the NLTK corpus. SNOMED-CT was tokenized and stop words were removed using the NLTK library. SNOMED-CT was found to contain 94,581 unique words and when combined with the Wordlist corpus, the union created a joint corpus of 304,760 unique words. This added 69,817 medical words facilitating more content for analysis. With only utilizing the Wordlist corpus, 6,008 words were matched from diagnosis records. The joint SNOMED-CT Wordlist corpus matched 6,535 words increasing the data size by 8.7 percent.

The severity score is calculated by determining the frequency distribution for all strings occurring in diagnosis records that resulted in survival and mortality. In N-gram language models, the probability of a sentence is computed by multiplying the frequencies of consecutive words. The number of words can vary, but usually a bigram or trigram model is chosen due to data sparsity:

$$p(s) = \prod_{i=1}^{i+1} p(w_i \mid w_{i-n+1}^{i-1}) \tag{6.1}$$

$$p(w_i \mid w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i})}{\sum_{w_i} c(w_{i-n+1}^{i})} \tag{6.2}$$

where n is the number of preceding words, $p(s)$ is the probability of a sentence, $p\left(w_i | w_{i-n+1}^{i-1}\right)$ is the probability of a word with n preceding words, $c\left(w_{i-n+1}^{i}\right)$ is the count of instances of w_i preceded by $w_{i-1}, w_{i-2}, \ldots, w_{i-n+1}$ words, $\sum_{w_i} c\left(w_{i-n+1}^{i}\right)$ is the count of all instances preceded by $w_{i-1}, w_{i-2}, \ldots, w_{i-n+1}$ words.

Unfortunately, the standard n-gram model will not accurately estimate sentence probabilities due to data sparsity. Smoothing techniques must be utilized or probabilities of sentences with zero count n-grams will also equal zero. Chen et al performed a comparative study of different smoothing techniques and found Modified Kneser-Ney Smoothing to have excellent performance [87]. Kneser-Ney Smoothing uses a single discount value D to reduce the counts of the n-grams, and then adds the counts of smaller n-1 grams. Modified Kneser-Ney Smoothing is a variation that uses three discount parameters as follows:

$$
\begin{aligned}
Y &= \frac{n_1}{n_1 + 2n_2} \\
D_1 &= 1 - 2Y\frac{n_2}{n_1} \\
D_2 &= 2 - 3Y\frac{n_3}{n_2} \\
D_{3+} &= 3 - 4Y\frac{n_4}{n_3} \\
D(c) &= \begin{cases} 0 & \text{if } c = 0 \\ D_1 & \text{if } c = 1 \\ D_2 & \text{if } c = 2 \\ D_{3+} & \text{if } c \geq 3 \end{cases}
\end{aligned}
\tag{6.3}
$$

where n_1, n_2, n_3, n_4 are counts for the number of n-grams with exactly 1, 2, 3 and 4 instances.

The following is the full equation for Modified Kneser-Ney Smoothing:

$$
p_{KN}(w_i \mid w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i}) - D(c(w_{i-n+1}^{i}))}{\sum_{w_i} c(w_{i-n+1}^{i})} + \gamma(w_{i-n+1}^{i-1}) p_{KN}(w_i \mid w_{i-n+2}^{i-1}) \tag{6.4}
$$

$$
\gamma(w_{i-n+1}^{i-1}) = \frac{D_1 N_1(w_{i-n+1}^{i}\bullet) + D_2 N_2(w_{i-n+1}^{i-1}\bullet) + D_{3+} N_{3+}(w_{i-n+1}^{i-1}\bullet)}{\sum_{w_i} c(w_{i-n+1}^{i})} \tag{6.5}
$$

where $N_1\left(w_{i-n+1}^{i-1}\bullet\right)$ is the number of n-grams that occur exactly once with the prefix words $w_{i-1}, w_{i-2}, \ldots, w_{i-n+1}$. N_2 and N_{3+} are defined analogously.

The next step is to calculate the posterior probability of patient mortality given the diagnosis sentence. The likelihood of the diagnosis sentence for survival and mortality are calculated using Modified Kneser-Ney smoothing. The priors for patient survival and mortality are determined from the dataset. The posterior probability of patient mortality is then calculated with Bayes formula:

$$
P(\pi_M \mid s) = \frac{P(s \mid \pi_M) P(\pi_M)}{\sum_{i \epsilon M,S} P(s \mid \pi_i) P(\pi_i)} \tag{6.6}
$$

where $P\left(s \mid \pi_M\right) P\left(\pi_M\right)$ and $P\left(s \mid \pi_S\right) P\left(\pi_S\right)$ are the likelihood of a sentence multiplied by the prior of patient survival or mortality.

After the posterior probability for mortality is calculated for every patient, they are clustered into 3 severity groups using K-means clustering. It was found that this approach accurately separated patients into groups with higher mortality rates (Table 6.5).

Table 6.5: Severity group results from K-Means clustering

C_s	*# Patients*	$\bar{P}(\pi_M \mid s)$	$P(M)$
1	577	0.0165	0.172
2	456	0.301	0.285
3	370	0.864	0.476

C_s severity group
$\bar{P}(\pi_M \mid s)$ average posterior probability for mortality in severity group.
$P(M)$ mortality rate in severity group.

Patients were then separated into 9 different groups based on high, medium and low severity and length of stay. The return rate and average return length of stay were calculated for each group and ICU. Since, there are five different ICU departments, it resulted in 45 groups of statistics (Table 6.6).

Table 6.6: Patient group results after clustering.

C_S	C_{LOS}	CV Surgery		Medical		Neurosurgery		Neuroscience		Surgical	
		$P(R)$	LOS_R	$P(R)$	LOS_R	$P(R)$	LOS_R	$P(R)$	LOS_R	$P(R)$	LOS_R
1	1	0.115	133.29	0.150	49.90	0.080	37.85	0.103	30.40	0.033	114.98
1	2	0.074	133.29	0.207	49.90	0.039	37.85	0.167	30.40	0.066	114.98
1	3	0.070	133.29	0.105	49.90	0.100	37.85	0.118	30.40	0.169	114.98
2	1	0.100	78.70	0.125	50.48	0.030	116.96	0.235	75.10	0.116	187.05
2	2	0.129	78.70	0.140	50.48	0.056	116.96	0.071	75.10	0.088	187.05
2	3	0.136	78.70	0.267	50.48	0.107	116.96	0.125	75.10	0.179	187.05
3	1	0.115	39.15	0.111	70.64	0.031	133.97	0.045	186.84	0.192	156.17
3	2	0.038	39.15	0.160	70.64	0.024	133.97	0.071	186.84	0.111	156.17
3	3	0.048	39.15	0.231	70.64	0.067	133.97	0.091	186.84	0.130	156.17

C_S Severity group

C_{LOS} LOS group

$P(R)$ Return rate for patient group

LOS_R Average return length of stay for patient group

6.3.3 Simulation Model

A simulation model was built using SimPy [348] to aid in the development and evaluation of the process flow of five intensive care units. A separate submodel was created for each ICU: CV Surgery, Neurosurgery, Medical, Neuroscience and Surgical. Each submodel had both scheduled and emergency arrivals. Scheduled arrivals were direct transfers after an appointed operation or surgical procedure, and emergency arrivals were unexpected admissions. The inter-arrival distributions were fitted using Matlab® 8.1 for both cases (Tables 6.2, 6.3).

Different numbers of beds were allocated and a separate queue was designated for each ICU. The full computer model contains scheduled and emergency arrivals for all five ICUs. Further, each ICU is modeled in detail, including service, queues, clinical and patient workflow, and their inter-dependencies on patient care and resources.

After a patient departs an intensive care unit, they are transferred to an intermediate care room before dismissal. The patient may return to an ICU after transfer to an intermediate room or they may exit the hospital.

Estimated probabilities from hospital records were utilized in the simulation model. The return and mortality rates were separately calculated for each ICU (Table 6.1). The Return module in our computerized model captures all possibilities for returns and exits. It also includes mortality cases where patients do not survive their ICU stay.

The simulation model tests six different queuing methods and each is executed in SimPy for a period of 7 days with ten replications. The results reported for each queuing model are averages over all replications.

6.3.4 ICU Resource Allocation

The goal of this system is to aggressively test the process flow of the hospital under heavy conditions. The given numbers of beds were approximated for each ICU according to an M/M/s queuing model. The model assumes there are s identical servers with unlimited waiting room capacity. Service duration follows an exponential distribution while arrivals occur at a constant rate according to a Poisson process. Given the number of servers s, average arrival rate λ, and average service time $\frac{1}{\mu}$, the mean waiting time in the queue W_q can be calculated under the M/M/s model by equation 6.10 [194]:

$$\begin{aligned} p_0 &= \left[\sum_{n=0}^{s-1} \frac{(\rho s)^n}{n!} + \frac{\rho^s s^{s+1}}{s!(s-\rho s)} \right]^{-1} \quad \rho < 1 \\ p_n &= \begin{cases} \dfrac{\lambda^n}{n!\ \mu^n}\, p_0 & (1 \leq n \leq s) \\ \dfrac{\lambda^n}{s^{n-s}\ s!\ \mu^n}\, p_0 & (n \geq s) \end{cases} \end{aligned}$$

$$p_D = 1 - \sum_{n=0}^{i-1} p_n \tag{6.7}$$

$$\rho = \lambda / s\mu \tag{6.8}$$

$$L_q = \frac{\rho}{1-\rho} p_D \tag{6.9}$$

$$W_q = L_q / \lambda \tag{6.10}$$

where equation 6.9 is the mean number of patients in the queue. Equation 6.7 calculates the probability that an arrival will experience a delay for service. Equation 6.8 is the average utilization for the queuing system, and s is the number of servers.

In the 2001 US National Hospital Ambulatory Medical Care Survey (NHAMCS), the average waiting time for an ICU bed reported was approximately 4.1 hours [321]. In this study, the average arrival rate and service duration were determined for each intensive care unit. Using the M/M/s model, the average wait times were calculated with the given number of beds for each ICU (Table 6.7).

Table 6.7: M/M/s Queuing Model parameters for each ICU.

ICU	λ	λ_s	μ	s_r	s_e
CV Surgery	0.125	0.013	0.011	18	16
Neurosurgery	0.105	0.034	0.015	20	11
Medical	0.122	0.002	0.017	14	11
Neuroscience	0.042	0.014	0.015	7	6
Surgical	0.140	0.032	0.013	20	15

$^{\lambda}$ Arrival rate (patients/hour)

$^{\lambda_s}$ Arrival rate from scheduled surgeries (patients/hour)

$^{\mu}$ Service rate (patients/hour)

^{s_r} Number of beds in hospital

^{s_e} Number of beds in simulation model

Parameters in the simulation model are determined empirically so as to match the hospital statistics for ICU admission delay to accurately evaluate the benefits for different test settings. Using the M/M/s model, performance measures were calculated for each ICU for different levels of bed availability. Since the M/M/s assumption of exponential service times can lead to underestimating actual congestion [192], the number of beds selected by the simulation model were associated with mean waiting times between 1.8 − 3.2 hours (Table 6.8): CV Surgery (16), Neurosurgery (11), Medical (11), Neuroscience (6), Surgical (15)

Table 6.8: Estimated wait times for each ICU using M/M/s Queuing Model.

CV Surgery		Neurosurgery		Medical		Neuroscience		Surgical	
11	705.541	8	52.162	8	60.122	3	267.681	11	413.788
12	51.580	9	14.948	9	15.611	4	23.198	12	42.430
13	18.190	10	5.684	10	5.834	5	5.606	13	15.265
14	7.972	11	2.333*	11	2.396*	6	1.531*	14	6.725
15	3.755	12	0.972	12	1.006	7	0.417**	15	3.170*
16	1.809*	13	0.400	13	0.419			16	1.525
17	0.871	14	0.160	14	0.171**			17	0.733
18	0.414**	15	0.062					18	0.348
		16	0.023					19	0.161
		17	0.008					20	0.073**
		18	0.003						
		19	0.001						
		20	0.000**						

s Number of beds

W_q average wait time (hours)

* W_q for s used by simulation model.

** W_q for s used by the hospital.

6.3.5 Classification of severity group

After a patient arrives at the hospital in the simulation model, they are classified into one of nine different groups based on their severity score and LOS. The LOS is generated from the distribution for the requested ICU. There are prior values for the percentage of patients in each severity group. However, the LOS distributions are slightly different for each severity group (Table 6.6). For example, it is rare to find a patient with high severity and high LOS. It would not be entirely accurate to assign the severity group based only on prior probabilities. Therefore, a posterior probability is calculated by multiplying the prior probability with the likelihood given a patient's LOS:

$$P(LOS; \mu_s, \sigma_s) = \frac{1}{LOS\ \sigma_s\ \sqrt{2\pi}}\ e^{-\dfrac{(\ln LOS - \mu_s)^2}{2\sigma_s^2}} \qquad (6.11)$$

$$P(LOS \mid C_s) = P(LOS; \mu_s, \sigma_s) \qquad (6.12)$$

$$P(LOS) = \sum_{s \in S} P(C_s) P(LOS \mid C_s) \qquad (6.13)$$

$$P(C_s \mid LOS) = \frac{P(C_s) P(LOS \mid C_s)}{P(LOS)} \qquad (6.14)$$

where C_s is the severity group class, LOS is the sampled value for length of stay from the ICU distribution, and μ_s and σ_s are parameters of the log-normal distribution for C_s in equation 6.11. Equation 6.12 is the likelihood of observing the LOS given C_s. Equation 6.13 marginalizes the overall probability for LOS over all severity cluster distributions. Equation 6.14 is the posterior probability of belonging to C_s given the LOS, and $P(C_s)$ is the prior probability of belonging to C_s,

The severity group is assigned to the admitted patient based on the calculated posterior probabilities for each class. Each group has a set of mortality rates determining whether the patient will die during their stay in the ICU (Table 6.5).

6.3.6 Managing Artificial Variability

There is substantial natural variability in hospital admissions through the emergency department, but there is also artificial variability. In this study, we found that 28.2% of entries were admitted to an ICU from elective surgeries. If adjusted for patient volume, scheduled surgical admissions can vary even more than through the Emergency Department (ED) [301]. This can have reciprocal effects where high surgical volumes can delay operations and increase waiting times for an available room. Operations can be cancelled due to a shortage of ICU beds.

In this study, the distribution is calculated for interarrival times to each ICU from scheduled and unscheduled admissions. A Passive model is first tested that uses no priority scheme and factors natural and artificial variability of arrivals. Each model reports the total patients served, severe patients admitted, average waiting times, return rate, and mortalities.

The Smooth Model is similar to the Passive Model, except it uses an ideal surgery schedule where there is no artificial variability. This is to help determine the effects the surgery schedule has on the hospital process flow. The average time between arrivals is calculated for scheduled admissions for each ICU (Table 6.8). Instead of using the fitted distributions for scheduled admissions, patients arrive at times equidistant from each other for each ICU.

The Smooth Model is not realistic, because even operation times can vary in ideal cases where elective surgeries are scheduled at efficient times. It is only used for evaluation purposes. All subsequent models utilize fitted distributions for scheduled admissions, but test different priority methods for admitting and bumping patients.

6.3.7 Priority Models

Typically, a queue admits entries on a first-come-first-serve (FCFS) basis. However, priority queues allow different classes to be treated differently. Without preemption, higher class items can jump ahead of others within the queue. However, service cannot be interrupted for any items in process. In a preemptive priority class, higher class items can discontinue other items currently in service [190]. In this study, both preemptive and non-preemptive models were tested to analyze the process flow of intensive care units.

Four different priority models were evaluated in our simulation model. Specifically, we derive and test models that both restrict and allow bumping while factoring the consequent mortality and return rates.

Greedy: The greedy method [459] gives patients with highest LOS the greatest priority. Using queuing theory, Siddhartan et al. showed that admitting patients with larger LOS before others lowered the overall average wait time [459]. The Greedy model is non-preemptive where bumping of patients is not permitted in any case. Higher priority patients in the queue are not permitted to interrupt lower priority patients in service.

Hybrid: The hybrid method admits patients based on their severity and their LOS. A patient in the highest severity group will be admitted first, but patients in the lower severity groups will be ordered according to their average LOS. The Hybrid model is also a non-preemptive method. It factors admission not only on efficiency, but also on the severity of the patients condition.

The next two priority models are both preemptive. They allow the service of lower priority patients to be interrupted if a higher priority patient is admitted.

Severity (Conservative) Bumping: The Conservative Bumping model is identical to the Hybrid model in the order patients are placed in the queue. However, a severe patient ($C_s = 3$) in the queue can bump a non-severe patient ($C_s < 3$) from service. Non-severe patients cannot bump any patients from service. Non-severe patients are bumped by lowest remaining length of stay plus the associated readmission load:

$$LOS_{tot} = LOS_{rem} + P(R_P) \times LOS_{Pret} \tag{6.15}$$

where LOS_{rem} is the remaining service time, $P(R_P)$ is the average return rate for the patient group, and LOS_{Pret} is the average service time for returns for the patient group (Table 6.6), LOS_{tot} is the estimated total service time. The readmission load is the product of return probability times return LOS, which is calculated using a similar method to the study by Chan et al [73].

Aggressive Bumping: Severe patients can still not be discharged from the ICU while in service. However, non-severe patients will be bumped when any type of patient requests admission to the ICU. Patients are discharged in the same order as the Conservative Bumping model. Aggressive Bumping is similar to the method used by Chan et al., except severe patients are restricted from ICU transfer before completion of service.

If a patient is bumped while in service, they will have a higher return rate as found in our hospital transfer records data (Table 6.4). Subsequently, the returned patients also have a higher mortality rate. All four different priority models are tested to determine the effects on waiting time, return rate and mortality.

6.4 Results

Table 6.9 reports the results for all six queuing models. Without enforcing any priorities for admission, the Passive Model reported higher average waiting time in the queue (62.03 min) and fewer total patients served (2,257).

The Smooth Model also does not enforce priorities, but arrivals from elective surgeries occur at a constant rate. The hospital only schedules surgeries Monday through Friday and operating hours can vary significantly. The Smooth Model is an ideal case that removes all variation from scheduled surgery arrivals. It gave impressive results when compared to the Passive Model at 56.22 minutes for average waiting time and 2,261 for total patients served. This raised the amount of patients as well as lowering delays. This showed reducing artificial variability is beneficial if it is possible to enforce a more regimented surgery schedule.

Table 6.9: Priority Queuing Model Results.

Model	# Patients	# Sev Patients	$P(R)$	$P(M)$	B	W_q	W_{qs}
Passive	2,257	547	0.089	0.286	0	62.03	63.57
Smooth	2,261	548	0.088	0.291	0	56.22	57.01
Greedy	2,306	559	0.091	0.286	0	36.22	28.86
Hybrid	2,287	564	0.084	0.288	0	35.86	13.20
Conservative Bumping	2,299	571	0.108	0.292	201	15.79	2.48
Aggressive Bumping	2,295	568	0.119	0.300	373	3.30	2.09

$^{P(R)}$ return rate
$^{P(M)}$ mortality probability
^{B} number of bumped patients
^{W_q} average waiting time in the queue for all patients
$^{W_{qs}}$ average waiting time in the queue for severe patients

Priority queuing models were tested with artificial variability utilizing the fitted distributions for scheduled surgery arrivals. The Greedy model only prioritizes patients by their expected LOS. It was able to serve 2,306 patients at an average waiting time of 36.22 minutes. This model could not capitalize on the benefits of uniform patient arrivals as with the Smooth Model, but it was able to significantly lower the average wait time due to more efficiently prioritization of patient admittance.

The Greedy model focuses on efficiency rather than patient severity. The Hybrid Model prioritizes severe patients above all others. Non-severe patients are prioritized by expected LOS identical to the Greedy Model. The Hybrid Model served 2,287 patients at 35.86 minutes average waiting time. Less patients are served than in the Greedy Model, but the average waiting time for severe patients was 13.20 minutes compared to 28.86 minutes. The Hybrid Model also had the lowest return rate at

8.4%.

The Conservative Bumping and Aggressive Bumping models reported results with substantial differences. Both preemptive queuing models prioritize patients by severity identical to the Hybrid Model. The Conservative model can only bump less severe ($C_s < 3$) patients from service when the most severe ($C_s = 3$) request ICU admission. The Aggressive Model bumps less severe patients from service for any patient requesting admission. The Conservative Bumping model served 2,299 patients and bumped an average of 201 from service. The average waiting time was 2.48 minutes for severe patients and 15.79 minutes for all patients. The mortality rate was only raised by 0.4 percent compared to the Hybrid Model. The Aggressive Model served 2,295 patients bumping 373 patients with an average waiting time of 3.3 minutes. The return rate increased by 1.0 percent and it reported the highest mortality rate for any model at 30.0%. Interestingly, it served four less patients than the Conservative Bumping Model and only 8 more patients than the Hybrid Model. It is clear that bumping can prove to be beneficial but only in heavily restricted cases.

6.5 Discussion

Healthcare centers that focus on operating at highest efficiency may consequently sacrifice the quality of care. By evaluating several different priority methods, the ICU system-based simulation model helps identify the costs of prioritizing by severity rather than efficiency. Severe priority methods do raise overall waiting times and lower the amount of patients served, but added benefits reduce further medical complications. Shorter wait times for severe patients result in lower return and mortality rates. Severe priority methods can show substantial enhancements by conservatively allowing bumping policies. Permitting early discharges with severe priority models resulted in wait times close to the most efficient models. However, without firm restrictions, bumping can significantly raise the mortality and return rates.

There are several potential future research studies that can be conducted with appropriate types of data. Our approach is applicable to other hospital data streams, for example, ICD diagnosis codes, patient resource needs, and hospital utilization status. Specifically, it would be beneficial to accurately categorize the diagnosis for each patient using individual ICD diagnosis codes. This would help determine if a patient return was due to an early discharge or because of an entirely new condition. Further, in our earlier readmission work [286], hospital resource usage and utilization information were employed to help predict patient readmission characteristics and the impact on patient needs and quality of care.

The studied hospital has five distinctly specialized intensive care units. There may be events where the requested ICU is full and a patient is diverted to an ICU of a different specialty [273]. It would be advantageous to examine the implications regarding permitted diversions for associated conditions. An analysis could be conducted whether patients benefit from diversions to ICUs of different specialties rather than remaining in the queue for the desired location.

Patient admissions can also be evaluated more globally. If estimated wait times were available for each hospital, the costs can be considered for redirecting patients to another hospital. The studied hospital herein has a sister medical center at a location about six miles away. It would be interesting to review records for cases where patients were blocked access and directed to this alternative location. A future study will analyze these cases and determine if transfer times were lower than estimated wait times for direct admission. Even in circumstances where total wait time were reduced by diversion, complications can result from the additional transit time. Optimizing patient flow in healthcare settings is a challenging balance between managing efficiency and maintaining quality of care. Hospitals can become more proficient and resourceful in daily operations by continuing to build system models that attempt to identify and investigate all significant interdependent factors.

CHAPTER VII

CLUSTERING PATIENTS BY SEMANTIC SIMILARITY

7.1 Introduction

Multimorbidity is the nonrandom association between 2 or more chronic medical conditions in one individual [508]. Patients with multimorbidities have a lowered quality of life with worsened medical conditions that are correlated with psychological distress and postoperative complications [153, 180, 527]. These factors result in higher costs of care in a healthcare industry [527] that is predominantly designed and evaluated by quality measurements for single diseased individuals [469, 470]. Multimorbidities are not rare isolated cases and can reach prevalence rates of over 50% in specific populations [154, 508]. According to a study by Barnett et al. [32], half of newborns are expected to suffer from multiple chronic conditions during their last 15 years of life. Investigative studies that facilitate a larger understanding of multimorbidity patterns will have profound implications for patient treatment decision support and prognosis strategies that focus care around the patient rather than the disease [227, 246].

Clinical practice guidelines are simply not relevant to "typical" patients with multiple chronic conditions [59, 306], because disease management and patient pathways are devised around single diseases [469, 470]. There is little evidence for patient treatment with multiple chronic conditions [316], and single-diseased guidelines often cannot address the complex needs of patients with multimorbidty [493]. Furthermore, physicians in many cases are unable to follow the best practices for each disease due to incompatible treatments [55, 408], such as interacting medications.

There is a need for strategies that can accurately identify sub-populations with multiple chronic conditions to develop treatment regimes that are relevant and appropriate for patients with multimorbidity. Several studies have focused on the highest absolute frequencies of specific combinations of diseases [513], but there are an enormous amount of theoretically possible permutations [509]. Given only a few diseases, there are already hundreds of thousands of different possible combinations requiring many calculations and very large samples. Newcomer et al. analyzed 17 conditions of interest and found only examples of 1,507 different combinations of coexisting conditions, but this is still too large for opportunities of targeted care management [355]. Frequencies of chronic conditions have limited value, since they are determined solely from prevalence rates in a population where the most common diseases return the most common combinations. It is more informative to view disease patterns from the nonrandom associations of health problems [507].

Data mining techniques provide an opportunity to identify meaningful groups of patients "up front" without calculating thousands of prevalence rates and conducting multivariable regression analyses. Previous studies have utilized multivariable regression analyses to predict patient outcomes [53,127,229], but it involves analyzing many patient-level characteristics on data sets that are often limited in size. Cluster analysis is a data mining technique that identifies groups of highly similar items within data sets [140]. It is exploratory by nature and has applications in many fields, such as market research, social network analysis, crime analysis, and climatology. Simple tabulations of disease frequencies are difficult to interpret, and clustering provides a unique opportunity to investigate the co-occurrence of multiple chronic conditions. By identifying specific combinations of comorbidities, healthcare outcomes can be improved from care management strategies that are more targeted to the individual.

7.2 Related Work

7.2.1 Semantic Similarity

When clinical documents are annotated to concepts in a standardized ontology, semantic similarity measures can be used to calculate the degree of relatedness between terms. Semantic similarity measures are particularly important to the biomedical domain, due to the large volumes of unstructured textual resources and the significance of terminologies. By classifying clinical records to standardized ontological concepts, patients can be semantically analyzed and subjects with similar conditions can be identified [432]. Information retrieval can be improved by extending user's queries with semantically similar terms [376]. Semantic technologies can also accelerate the integration of heterogeneous clinical information by disambiguating medical concepts and identifying equivalencies when linking separate sources [478]. Ontologies have become an important component of biomedical research, and in response, a wide variety of semantic similarity approaches have been developed. However, there is no clear choice for the best strategy, because different measures perform differently depending on the circumstances. Results can vary depending on the corpus, benchmark, or applied terminology [173, 385].

As discussed in Section 2.3, semantic similarity approaches can be divided into two broad categories, edge-based measures and information content measures. Edge-based measures simply compare entities by counting the amount of taxonomic links between items. Rada et al. [403] calculates the distance between items as the minimum path connecting them:

$$dis_{rad}(c_1, c_2) = min_{\forall i} \left|path_i(c_1, c_2)\right| \tag{7.1}$$

Wu and Palmer [533] calculate the minimum path but also factor the relative depth of both concepts in the taxonomy. The similarity measure is determined from the depth of the Least Common Subsumer (LCS), the most specific ancestor common to both concepts:

$$sim_{wp}(c_1, c_2) = \frac{2 \times depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \tag{7.2}$$

Leacock and Chodorow [282] use the minimum path and maximum taxonomy depth to calculate similarity:

$$sim_{lch}(c_1, c_2) = -log\left(\frac{\delta(c_1, c_2)}{2 \cdot max\_depth}\right) \tag{7.3}$$

where $\delta(c_1, c_2) = min_{\forall i} |path_i(c_1, c_2)|$.

Information content (IC) measures compute similarity by determining the amount of shared information between concepts. IC quantifies the amount of information a term provides. The IC of a term is calculated by:

$$IC_{corpus}(c) = -log(p(c)) = -log(\frac{freq(c)}{freq(root)}) \tag{7.4}$$

where $p(c)$ is the probability of a concept c appearing in a corpus. Resnick [411] quantified similarity by the IC of the most informative common ancestor (MICA):

$$sim_{res}(c_1, c_2) = IC(MICA(c_1, c_2)) \tag{7.5}$$

The Lin [297] and Jiang & Conrath [240] measures also factor the IC of the compared concepts:

$$sim_{lin}(c_1, c_2) = \frac{2 \cdot IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{7.6}$$

$$dist_{jcn}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \cdot IC(MICA(c_1, c_2)) \tag{7.7}$$

Resnick, Lin, and Jiang & Conrath are all corpus-based measures, because concept appearance frequencies must be estimated from the given corpora. Textual corpora contain words that must be disambiguated and matched to standardized concepts in the medical ontology. The probability of each concept may not be calculated accurately due to the sparseness of the data, which is especially evident in the clinical domain, due to limited coverage of biomedical terms [376].

To overcome the limitations of corpus-based IC measures, the principles of information theory are applied in recent studies to create an intrinsic IC computation. The following measures do not rely on corpora, but are more advanced than edge-based measures by incorporating additional semantic evidence. Seco et al. [450], Zhou et al. [547], and Sanchez et al. [433] depend on values, such as the number of hyponyms, subsumers, and leaves:

$$IC_{seco}(c) = 1 - \frac{log(hypo(c)+1)}{log(hypo(root))} \tag{7.8}$$

$$IC_{zhou}(c) = k(IC_{seco}) + (1-k)(\frac{log(depth(c))}{log(max\_depth)} \tag{7.9}$$

$$IC_{sanchez} = -log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \tag{7.10}$$

Sanchez et al. [433] reported that its metric outperformed both intrinsic IC calculations by Seco et al. and Zhou et al. when comparing correlations against the Miller and Charles benchmark [338]. A subsequent study by Sanchez et al. [432] evaluated semantic similarity correlation with the Pedersen benchmark [376] using the SNOMED CT ontology and the Mayo Clinic Corpus of Clinical Notes [376]. Results were found that almost all measures based on intrinsic IC calculations outperformed corpus-based measures.

With effective patient distance metrics based on semantic similarity, healthcare can be expanded by accurately identifying relevant cases of interest and producing medically relevant interpretations. Melton et al. [331] evaluated five distance metrics to determine patient similarity. The metrics varied on how deeply the ontology or information content was incorporated. Patient records were observed by experts who manually assessed similarity scores. An expert correlation was calculated for each distance metric. The author concluded that information content and the usage of an ontology improve biomedical semantic similarity measures but fall far short of performance compared to experts. Cao et al. [62] advance patient similarity by only extracting features pertaining to disease, finding, medication, and procedure. Patient similarity is calculated by determining the probability of each feature in the given corpus and using the Aslam and Frost metric [21]. Plaza et al. [392] map terms in patient records to UMLS and create concept graphs. Concepts in each graph are weighted based on how highly they are connected in the taxonomy. Similarity is then calculated by comparing each patient's concept graph. Pivovarov and Elhadad [391] introduce a hybrid-based approach by combining ontological and corpus based methods and determining similarity by context vectors and SNOMED ontological knowledge. Mabotuwana et al. [308] extract concepts from medical reports into a document vector space model but expanded to a semantic vector [312]. Each feature in the semantic vector represents a concept in the SNOMED ontology. For all mapped concepts from a patient records, feature values are set to 1. All parent concepts are set proportionally smaller based on their distance from the mapped concepts. Chan et al. [74] received positive performance by using a simpler vector model. Using an ontological vector, features were set to only a certain depth in the hierarchy. It was found that level 4 returned the best similarity results when matching a query to database health records. This study also uses an ontological vector and SNOMED CT to calculate patient similarity, but unlike Chan et al., it investigates several semantic similarity measures. Chan et al. and Mabotuwana et al. both use a a simple inverse of path length to calculate the similarity to higher level concepts. In this study, the semantic similarity measure with the highest performance is used to set the feature value of level 4 concepts.

7.2.2 Clustering

Cluster analyses has previously been used in health care services research to discover patterns of multimorbidities, but it is an exploratory classification method where different algorithms can produce difference results [140]. There is a large variety of clustering methods, and a specific algorithm may be more appropriate than others depending on the dataset. The most common distinction between clustering types is whether clusters are hierarchical (nested) or partitional (unnested). In hierarchical clustering, clusters are permitted to have subclusters and the set of nested clusters are organized as a tree. In partitional clustering, the entire dataset is divided into non-overlapping clusters where items belong to only one subset. Different clustering algorithms are defined by the specific procedures and steps that are followed to determine each subset of data. However, clustering may also be categorized by the methods that objects can be assigned to each cluster. Items may belong to clusters exclusively or assignments can be overlapping. Overlapping clustering allow objects to simultaneously belong to more than one group. In fuzzy clustering, items belong to every cluster, but a probability is calculated to determine the weight of each membership to each group. Additionally, clustering may be complete or partial. Complete clustering assigns every object in the data set to a cluster while in partial clustering this rule is not enforced. Certain items may be outliers or noise, and partial clustering is advantageous in these cases where it is not preferable to assign clusters to the complete dataset [485].

As there are an array of different clustering algorithms and item assignment methods, there is also variation in the types of clusters that are generated. In Prototype-based clusters, all objects are closer to the prototype of their assigned cluster than to the prototypes of the remaining clusters. A prototype is typically the average (mean) of all the points in the cluster. K-means is a prototype-based partitional clustering technique [310]. In Graph-based clusters, a group of objects are connected to only to one another but not to any objects outside the group. Agglomerative Hierarchical Clustering [351] can be interpreted as graph-based clustering by repeatedly merging connected components. Density-based clusters form dense regions of objects while ignoring noisy objects in low density regions. DBSCAN [137] is a density-based clustering algorithm that produces a partial clustering of the dataset by ignoring noise and outliers.

The accurate use of cluster analysis to identify clinically relevant comorbidity patterns depends on selection of a suitable algorithm with appropriate distance measures that can quantify degrees of nonrandomness. In previous studies, there is high heterogeneity in the applied methodological criteria for knowledge discovery of patients with multiple chronic conditions [395]. This leads to a diversity of multimorbidty patterns found across many different studies. Exploratory factor analysis [346] is a common alternative technique to clustering to identify symptom groups [172,219,260, 393,396,444]. Its objective is to map high numbers of correlated observed variables into a common causal factor. The main motivation for exploratory factor analyses are association measures that adjust for non-random comorbidity and disease cluster visualizations with health problems that overlap into multiple groups [393,396]. However, exploratory factor analysis requires data to be in a continuous format and authors have utilized tetrachoric correlations to bypass this restriction for diagnostic variables [275,395]. Also, factor models may fail to correspond to the common causes of symptoms where methods that permit a wider variety of causal specifications can be more successful [363]. Multiple correspondence analyses [171], Unified clustering [357], and Structural Equation Modelling [363] (SEM) are less conventional implementations that attempt to accommodate for the limitations of exploratory factor analysis and clustering.

Clustering is the most common technique used in identifying multimorbidity patterns, due to its ability to isolate potentially meaningful disease clusters without initially specifying all theoretically possible combinations. Several studies group diseases by hierarchical clustering and utilizing a Jaccard coefficient or Yule Q as a proximity measure [106,243,318,319]. However, these techniques do not allow health problems to simultaneously exist in multiple clusters. By hierarchically clustering patients instead of diseases, Newcomer et al. [355] and Goldstein et al. [185] both bypass this limitation. Despite successfully returning multimorbidity patterns, hierarchically clustering is not without its technical difficulties. They have a lack of robustness, each data point is forced into a cluster, and dendograms are a complicated interpretation of determining the appropriate number of clusters [344]. The current study utilizes a density-based clustering algorithm, known as OPTICS [11] with a semantic distance proximity measure that alleviates many of these shortcomings. Clusters vary by density relieving the necessity to initiate parameters that significantly influence results. Patients are not coerced into cluster assignment by ignoring outliers and focus is placed only on similarly related items.

7.3 Methods

Using the techniques described in Section 5.3, the diagnoses and procedures from ICU stays for all patients from the MIMIC-II database were mapped to SNOMED CT concepts. Similar to the study by Chan et al. [74], age, gender and associated SNOMED CT concepts were mapped to an ontological vector:

$$v = [g_1 g_2 e_1 e_2 \dots e_{100} a_1 a_2 \dots a_m]^T \tag{7.11}$$

where g_1 and g_2 are Boolean variables for male and female, e_k is a Boolean variable for age $k \in [1, 99]$ and e_{100} for $age \geq 100$. a_i represents the patient record's semantic similarity to the ith SNOMED CT level 4 concept where $i \in [1, m]$ and m is the amount of concepts in the SNOMED CT taxonomy at depth level 4. Level 4 nodes have a shortest path of 4 edges to the head node.

Chan et al. maps all SNOMED CT concepts from a patient EHR to their equivalent level 4 nodes in the SNOMED CT hierarchy. In the ontological vector, each element a_i is set to the maximum semantic similarity between the ith level 4 node and all descending SNOMED CT concepts identified in the patient EHR. If the level

4 node has no concept descendants in the EHR, a_i is set to 0. Chan et al. calculate semantic similarity with the following equation:

$$\frac{1}{1 + min_{j=1..n} d_{ij}} \tag{7.12}$$

where d_{ij} is the shortest path length from the ith SNOMED CT level 4 concept and the jth SNOMED CT patient EHR concept.

This is similar to the path length similarity measure defined by Pedersen et al. [376]. However, in analyzing semantic similarity measures, it was found in previous studies that path length has low correlation with experts [376]. There are many biomedical semantic similarity measures that are divided into higher level approaches that are either based on edge counting or information content (See Section 2.3). Different measures perform differently, and results depend on the dataset and field of interest in the biomedical domain [173, 385]. Due to these discrepancies, it is important to test the most popular methods with the current MIMIC-II dataset developed in Section 5.3. There are nearly 350,000 nodes in the SNOMED CT Graph Database, and it is apparent that semantic similarity cannot be calculated between all possible node pairs. However, the studies by Pedersen et al. [376] and Hliaoutakis [217] both provide a list of biomedical concept pairs and similarity scores manually assigned by experts. These benchmarks have been used to evaluate semantic similarity measures in multiple studies [6, 37, 371, 376, 432, 433].

In this study, nine different measures are calculated for semantic similarity between SNOMED CT concepts. These include edge-based measures (Path [376], LCH [282] and Wu & Palmer [533]), corpus-based information content (IC) measures (Resnick [411], Lin [297], Jiang & Conrath [240]), and intrinsic IC measures (Seco [450], Zhou [547] and Sanchez [433]).

For information content measures, counts are calculated for all mapped SNOMED CT concept to ICU patient stays. These counts are then assigned as properties to the equivalent nodes in the Neo4j SNOMED CT Graph Database. For information content, a parent node inherits the counts of all its children nodes. Therefore, after assignment of the corpus counts to the current node, the counts are percolated upward and added to the counts of all parent nodes. In this fashion, the head node will always be the node with the highest count.

Almost all semantic similarity measures require identification of the lowest common ancestor between compared concepts. A Cypher query is constructed to retrieve the lowest common ancestor for all concept pairs from the Pedersen et al. [376] and Hliaoutakis [217] benchmarks. After completion of percolating term counts and graph queries, all required information is provided to calculate each semantic similarity measure. The semantic similarity measure with the highest correlation is used for the value of the level 4 node in the ontological vector. Correlation is calculated by Spearman's rank correlation coefficient [289]:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{7.13}$$

where d_i is the difference between ranks. The Sanchez intrinsic IC measure is found to have the highest correlation and is more effective than the Path similarity measure used in the study by Chan et al. [74]. Path was reported as one of the lowest performing metrics when compared to other semantic similarity measures [376, 387].

After the ontological vector for each patient ICU stay is calculated, a similarity matrix is constructed with each entry containing the cosine distance between vectors:

$$1 - \frac{\vec{u} \cdot \vec{v}}{||\vec{u}|| ||\vec{v}||} \tag{7.14}$$

Since an adjacency matrix is symmetric, it is only necessary to perform $\frac{n(n-1)}{2}$ instead of n^2 operations. With a matrix containing an entry for all possible patient comparisons, it is possible to create similarly related groups from the many available clustering procedures. A density-based clustering procedure is selected due to its ability to discard outliers and infer the number of clusters. DBSCAN is a density-based clustering algorithm that is able to discover clusters of arbitrary shape and handle noise and outliers effectively [137]. For each cluster, the neighborhood of a given radius (Eps) must contain the minimum number of members (MinPts). To aid in the appropriate selection of Eps, a k-distance distribution can be generated for the dataset [435]. By setting k to the MinPts in each cluster, the maximum Eps distance is calculated for each point's k-nearest neighbors. A distribution of these k-distances (k=30) is then created for all patient ICU stays in the MIMIC-II database (Figure 7.1). K is set to 30 as an appropriate minimum number of patients in each cluster. In observing the k-distance distribution, the distribution substantially increases at a threshold value of roughly 0.30 for epsilon.

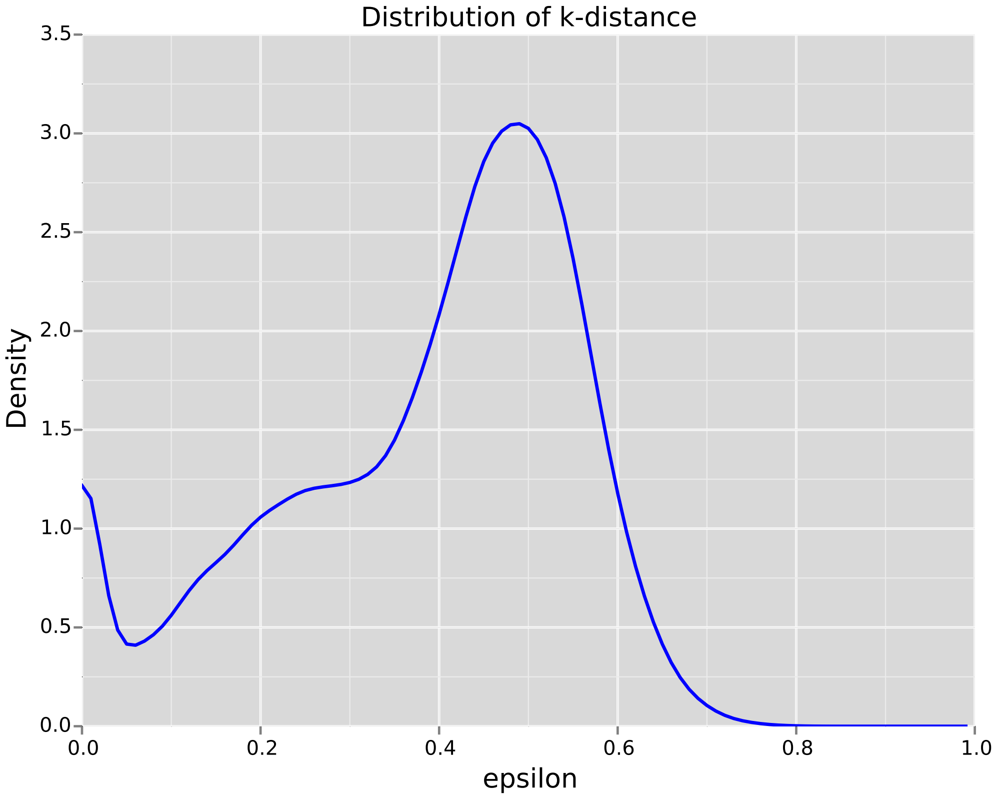


Figure 7.1: K-distance distribution of MIMIC-II patient ICU stays (K=30).

In the dataset, it is very possible that there may be clusters of patients that are moderately separated while others are more densely packed. However, in DBSCAN, epsilon is a set value and the expectation is that clusters will have approximately the same density. In response to the limitations of DBSCAN, the OPTICS [11] algorithm was developed to allow clustering of varying density. The output of the algorithm is a reachability plot which is a bar plot of distances, and clusters are represented by "dents" in the plot. However, it is not explicit how to determine the clusters from the reachability plot, and visual and automatic methods can vary without a clear best choice for a horizontal cut-line. Sander et al. [436] developed an algorithm to automatically create a hierarchical representation of clusters from a reachability plot by setting MinPts $= 2$ and Eps $= \infty$. Under these settings, the core-distance of every object is its distance to its nearest neighbor. The methods in this study use the OPTICS algorithm with the Sander et al. settings, and new clusters are created each time the reachability plot increase by a distance of 10% of the average of its leftmost and rightmost neighboring points. If the containing points of a newly formed cluster already exist in smaller clusters, the smaller clusters will become hierarchical children of the new larger cluster.

Hierarchical clustering with OPTICS does return a superfluous amount of clusters, because there is a great deal of overlapping where more specific clusters are contained within larger clusters. To alleviate this issue, the average distance between all cluster members are calculated. Any cluster is removed that exceeds an average distance of 0.3. In the groups leftover, if any larger cluster contains smaller clusters, the smaller clusters are disregarded. Finally, all clusters where 10% of patients do not contain a morbidity condition are not included. This yields a much smaller and more informative list of clusters of multimorbidity patients than the initial groups originally returned by the OPTICS algorithm.

7.4 Results

The SNOMED CT Graph Database built by Neo4j was utilized to calculate the lowest common ancestor between all compared concepts in the Pedersen et al. [376] and Hliaoutakis [217] benchmarks. A total of 60 common ancestors were computed in 156.13 seconds. This was a significantly lightweight implementation when compared against popular semantic web [43] applications that utilize Triple Stores. Triple Stores contain data entities that are only composed of subject-predicate-object fields. They can be stored in Resource Description Framework (RDF) [337] or Web Ontology Language (OWL) [328], and SPARQL [379] is the standard web query language for RDF. SNOMED CT can be converted into OWL format using the OWL transform script provided in the downloaded SNOMED CT Release files. However, after conversion, the loading of the OWL for SPARQL processing completed in 448.3 seconds and required up to 9 Gb in memory. Neo4j contained an optimized data structure for graph networks and was a much more suitable choice for traversing large ontological networks.

Of the nine separate semantic similarity measures tested, all edge counting measures performed the poorest. Generally the Pedersen [376] benchmark gave higher values for correlation than Hliaoutakis [217]. However, no edge counting measure exceeded a correlation value of 0.7 for the Pedersen benchmark. For information content measures, despite its simplicity, Resnik scored the highest in both studies. The intrinsic Seco and Zhou scores both did not perform as well as the corpus based information content measures. However, the most up to date intrinsic IC measure by Sanchez et al. returned the highest average of both studies. Therefore, the intrinsic IC Sanchez measure was used for calculation in creating the ontological vector of semantic similarity to level 4 SNOMED CT concepts. All results can be viewed in Table 7.1.

Table 7.1: Semantic similarity correlation with experts

Measure	*Pedersen 2007* [376]	*Hliaoutakis 2005* [217]
Path [376]	0.694	0.335
LCH [282]	0.694	0.335
Wu & Palmer [533]	0.554	0.318
Resnik [411]	0.808	0.49
Lin [297]	0.796	0.474
Jiang & Conrath [240]	0.754	0.482
Intrinsic (Seco) [450]	0.65	0.486
Intrinsic (Zhou) [547]	0.754	0.469
Intrinsic (Sanchez) [433]	0.754	0.535

After creation of the ontological vector for each ICU stay, the OPTICS clustering algorithm executed in 2.48 hours. The graph of reachability distances can viewed in Figure 7.2. As observable in the graph, there are many valleys where patients become more densely clustered.

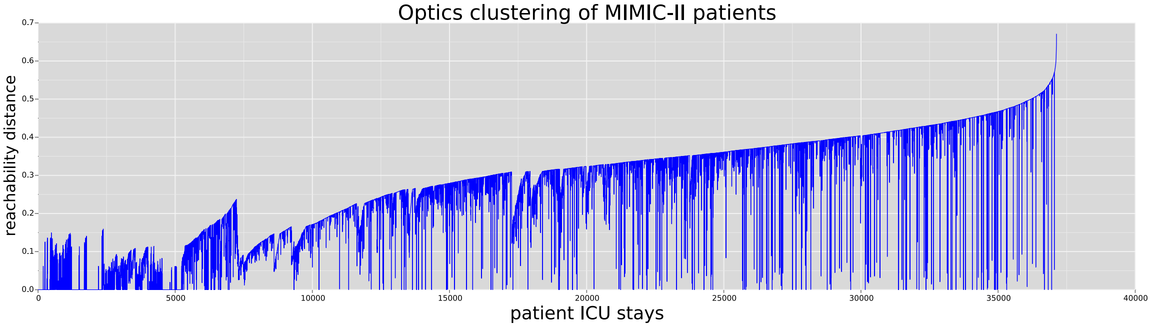


Figure 7.2: Reachability Plot from OPTICS clustering of semantic distances between MIMIC-II patient ICU stays.

Hierarchical clustering with optics does return a superfluous amount of clusters, because there is a great deal of overlapping where more specific clusters are contained within larger clusters. To alleviate this issue, the average distance between all cluster members was calculated. Any cluster was removed that exceeded an average distance of 0.3. In the groups leftover, If any larger cluster contained smaller clusters, the smaller clusters were disregarded. Finally, all clusters where 10% of patients did not contain a morbidity condition were not included. This yielded a much smaller and more informative list of clusters of multimorbidity patients than the initial groups originally returned by the OPTICS algorithm.

The hierarchical clustering algorithm is performed on the set of reachability distances in 2.58 minutes and over 2,331 clusters are returned. This reduces to 449 after removing clusters with a greater average distance than 0.3. 332 smaller clusters were then eliminated due to already being represented in larger clusters. The final clusters were excluded if 10% of patients in the group did not contain a morbidity. As a result, 22 multimorbidity clusters totaling 1,030 patients were identified in conclusion of the process. The general conditions of all patients in each cluster were determined by mapping their ICD-9 codes into ISHMT format. The procedures with the largest frequency were also calculated. The statistics for each cluster can be viewed in Table 7.2.

Each cluster is identified by the most frequent diagnoses and procedures of all included patients. Prados-Torres et al. [395] surveyed 14 multimorbidity studies and found 97 separate disease patterns were reported primarily due to the variation in approaches and datasets. However, the three most common themes were metabolic syndromes (heart disease, diabetes, and obesity), mental health, and musculoskeletal disorders. The metabolic syndrome is consistent in the clusters identified in our study. Clusters 3, 7, 8, 9 are all comprised of conditions related to cardiovascular disease and metabolic disorders. The clusters can vary by the specific diagnoses, such as coronary arteriosclerosis, angina, high cholesterol, or diabetes. However, the clusters also differ between the common procedures. This helps understand the prevalent operations for each group of patients, and at a higher level of granularity, can potentially help determine the best treatment for clusters with similar conditions but different procedures.

Cardiopulmonary bypass operation has the highest frequency in clusters 7-9, but angiocardiography and placement of stent are the most common procedures in cluster 3. Metabolic disorders have been found to be a precursor to Cardiovascular disease and Type 2 Diabetes [526], and similar metabolic clusters were identified in previous studies [106, 260, 396, 444]. High cholesterol is correlated to cerebrovascular disease both in clusters 11 and 14, and it has been shown that cholesterol can be an important precursor to stroke risk [211]. Interestingly, cluster 14 also has patients with depressive disorder, and depression is frequently described as a condition that follows stroke [201]. A similar cluster was found by [444]. Cluster 15 is diabetic neuropathy where long-standing diabetes can develop nerve damage throughout the body that affects the digestive and renal system. Cluster 16 is diabetic ketoacidosis where burning fatty acids produce ketone bodies that result in excessive urine production and potentially impaired kidney function due to dehydration. Clusters 12 and 21 contain patients with peptic ulcers, anaemias, and gastroduodenitis. Anemia is one of the symptoms of an ulcer. Helicobacter pylori is a bacteria that infects the stomach and its responsibility in causing gastroduodenitis and peptic ulcer development has been studied [538]. A similar cluster was identified in the longitudinal analysis by Wong et al. [528].

Clusters 2, 4, 5, and 6 all cover neonatal patients. More neonatal clusters are developed than previous multimorbidity works that have higher focus on elderly populations [55, 152, 318, 319, 393, 444, 527], and the current clustering method also calculates patient separation by age and gender. Most neonatal clusters contain neonatal jaundice, and it can be caused by a multitude of factors, such as respiratory distress, sepsis, and anemia [394]. Cluster 19 covers intestinal diseases and several neoplasms, including ovarian and cervical cancer. One of the most common malignancies that cause bowel obstruction is ovarian cancer [415]. Kirchberger et al. also found a similar cluster with gastrointestinal disorders and cancer [260]. Cluster 22 includes biliary tract disorders and cancers. Cluster 20 comprises of patient with neoplasms but with kidney cancer and diabetes. Associations between these two conditions have been investigated [200]. Cluster 13 covers patients with respiratory failure where one of the causes may be pneumonia, and urinary tract infections and cardiac arrythmias are possible subsequent complications [255]. As each cluster is scientifically verified, it is apparent that not all conditions follow a strict causal relationship, highlighting one of the main disadvantages of previous studies that utilize exploratory factor analysis [172, 219, 260, 393, 396, 444].

Table 7.2: Descriptions of patient clusters after optics clustering

Cluster	*Description*	*Patients*	*Median Age* (25th Percentile, 75th Percentile)	*Most Prevalent Conditions and Procedures*
1	Cardiovascular disease with diabetes and stomach disease	171	62.4 (54.4,69.3)	Coronary arteriosclerosis 100.00% Hypercholesterolemia 55.56% Diabetes 35.67% Gastroesophageal reflux disease 17.54% Atrial fibrillation 16.96% Hyperlipidemia 15.79% Tobacco dependence 8.77% Catheterization of left heart 94.74% Cardiopulmonary bypass operation 94.15% Angiocardiography of left heart 77.78%
2	Neonatal cardiovascular disease with respiratory disease and jaundice	59	Neonate	Patent ductus arteriosus 100.00% Respiratory distress syndrome in the newborn 93.22% Neonatal jaundice 98.31% Neonatal bradycardia 33.90% Anemia of prematurity 30.51% Phototherapy 94.92% Insertion of endotracheal tube 77.97% Continuous positive airway pressure 71.19% Enteral feeding 72.88%
3	Cardiovascular disease with high cholesterol and elevated lipid levels	144	56.8 (52.1,64.4)	Coronary arteriosclerosis 100.00% Hypercholesterolemia 35.42% Hyperlipidemia 14.58% Diabetes Mellitus 14.58% Injection of therapeutic agent 100.69% Placement of stent in coronary artery 50.69% Angiocardiography of left heart 32.64% Angiocardiography of right heart 23.61%
4	Neonatal jaundice with respiratory disease	28	Neonate	Neonatal jaundice 100.00% Respiratory distress syndrome in newborn 25.00% Neonatal bradycardia 82.14% Phototherapy of newborn 82.14% Enteral feeding 57.14% Continuous positive airway pressure 25.00%

5	Neonatal jaundice with respiratory disease and sepsis	22	Neonate	Neonatal jaundice 100.00% Patent ductus arteriosus 86.36% Bacterial sepsis of newborn 77.27% Chronic respiratory disease in perinatal period 50.00% Anemia of prematurity 50.00% Neonatal bradycardia 45.45% Retinopathy of prematurity 40.91% Phototherapy of newborn 90.91% Enteral feeding 90.91% Diagnostic lumbar puncture 86.36% Artificial respiration 86.36% Insertion of endotracheal tube 86.36%
6	Neonatal conjunctivitis with jaundice and respiratory disease	46	Neonate	Neonatal conjunctivitis 100.00% Neonatal jaundice 78.26% Respiratory distress syndrome 43.48% Neonatal bradycardia 32.61% Phototherapy of newborn 78.26% Enteral feeding 76.09% Artificial respiration 39.13% Insertion of endotracheal tube 39.13%
7	Cardiovascular disease with infarction, diabetes mellitus, and high cholesterol.	47	66.1 (57.6,70.8)	Impending infarction 100.00% Coronoary arteriosclerosis 100.00% Hypercholesterolemia 61.70% Diabetes Mellitus 48.94% Atrial fibrillation 21.28% Hyperlipidemia 19.15% Cardiopulmonary bypass operation 95.74%
8	Cardiovascular disease with angina, elevated cholesterol, and diabetes mellitus	41	64.5 (60.2,67.8)	Coronoary arteriosclerosis 100.00% Diabetes Mellitus 65.85% Hypercholesterolemia 82.93% Angina 34.15% Cardiopulmonary bypass operation 90.24%
9	Cardiovascular disease and surgical complications with elevated cholesterol and diabetes mellitus	38	69.3 (59.0,72.2)	Coronary arteriosclerosis 100.00% Hypercholesterolemia 73.68% Postoperative cardiac complication 50.00% Atrial fibrillation 50.00% Diabetes Mellitus 28.95% Angina 23.68% Cardiopulmonary bypass operation 97.37%
10	Cardiovascular disease with aortic insufficiency and gastroesophageal reflux disease	36	52.2 (40.1,59.7)	Aortic insufficiency 71.05% Congenital aortic insufficiency 55.26% Gastroesophageal reflux disease 15.79% Hyperlipidemia 13.16%
11	Cerebrovascular disease with high cholesterol	45	56.9 (44.2,60.3)	Cerebral Aneurysm 53.33% Cerebral hemorrhage 37.78% Hypercholesterolemia 13.33% Arteriography of cerebral arteries 100.00% Aneurysm clipping 37.78% Embolization 22.22%
12	Digestive disease with peptic ulcer, anaemias, and cardiac arrhythmias	30	72.5 (68.2,81.2)	Peptic ulcer 80.00% Anaemias 73.33% Atrial fibrillation 23.33% Heart failure 23.33% Hypothroidism 16.67% Hypercholesterolmeia 16.67% Endoscopic control of duodenal bleeding 56.67% Endoscopic control of gastric bleeding 43.33% Esophagogastroduodenoscopy 40.00%
13	Respiratory disease with pneumonia, cardiac arrythmias, and urinary system disorders	25	77.8 (69.5,88.2)	Acute respiratory failure 56.00% Pneumonia 52.00% Atrial fibrillation 48.00% Urinary tract infectious disease 36.00% Artificial respiration 80.00% Tracheostomy 72.00% Insertion of endotracheal tube 72.00% Enteral feeding 48.00%

14	Intracranial injury with cardiac arrhythmias, and high cholesterol.	42	79.7 (72.0,83.9)	Subdural hemorrhage 100.00% Atrial fibrillation 26.19% Hypercholesterolemia 19.05% Coronary arteriosclerosis 16.67% Incision of cerebral meninges 61.90%
15	Diabetes mellitus with nerve disorders, arteriosclerosis, and kidney disease, stomach disease	21	37.2 (35.3,37.9)	Diabetic autonomic neuropathy 100.00% Gastroparesis 100.00% Neurological disorder 80.95% Coronary arteriosclerosis 90.48% Renal Failure 52.38%
16	Diabetes Mellitus with metabolic disorders and kidney disease	34	22.1 (20.4,31.4)	Ketoacidosis 100.00% Diabetes Mellitus 100.00% Hyperlipidemia 35.29% Depressive disorder 26.47% Renal Failure 26.47%
17	Fractures and respiratory injury	48	42.8 (26.5,50.9)	Fracture 100.00% Pulmonary contusion 35.42% Traumatic Pneumothorax 33.33%
18	Esophagus disorder with alcoholism, cirrhosis, and platelet disorders	38	52.5 (45.6,55.0)	Esophageal varices 73.68% Alcoholic cirrhosis 60.53% Cirrhosis of liver 34.21% Chronic hepatitis C 55.26% Alcoholism 52.63% Thrombocytopenic disorder 34.21% Esophagogastroduodenoscopy 26.32% Esophageal polypectomy 23.68% Platelet Transfusion 23.68% Esophagoscopy 21.05%
19	Gynaecological cancer with intestinal disease	24	58.1 (48.2,70.8)	Peritoneal cancer 100.00% Ovarian cancer 83.33% Cervical cancer 62.50% Paralytic ileus 33.33% Intestinal cancer 25.00% Bilateral salpingectomy with oophorectomy 79.17% Excision of lesion of mesentery 75.00% Supracervical hysterectomy 25.00%
20	Kidney Cancer with diabetes mellitus and anaemias	21	66.1 (53.3,78.5)	Kidney Cancer 100.00% Diabetes mellitus 38.10% Anaemias 33.33% Nephroureterectomy 85.71% Excision of regional lymph nodes 33.33%
21	Digestive disease with peptic ulcer and anaemias	24	59.3 (46.6,75.4)	Peptic ulcer 91.67% Anaemias 83.33% Hypertension 29.17% Gastroduodenitis 41.67% Packed blood cell transfusion 83.33% Esophagogastroduodenoscopy 83.33%
22	Hepato-biliary and pancreas diseases and neoplasms	46	(51.8,82.4)	Gall Bladder and Biliary Tract Disease 76.09% Diseases of Pancreas 39.13% Neoplasms of liver, pancreas or bile duct 34.78% Endoscopic retrograde cholangiopancreatography 100.00% Artificial respiration 30.43% Insertion of endotracheal tube 23.91%

7.5 Discussion

Results from previous multimorbidity clustering studies have greatly varied mainly due to the high heterogeneity in the chosen methodology. Approaches have differed from sample size, number of baselines diseases, age recruitment, and the chosen statistical procedures [395]. In most cases, the chronic diseases used for analysis of multimorbidities were chosen beforehand and decided by prevalence rates [318, 319, 395, 444, 510]. In this study, since all diagnoses are mapped to standardized concepts in SNOMED CT, it is not necessary to make prior decisions of which conditions to exclude or include in the clustering of patients. Prior methods that simply report the most common co-occurring conditions have proven unsuccessful, because they do not accurately account for nonrandom association nor are they influenced by the complete case history of the patient. Cluster analysis methods have sufficiently identified multimorbidity groups with validated clinical basis from previous research [106, 243, 318, 319]. However, the hierarchical techniques of agglomerative clustering create a complicated denndrogram that is not easily interpretable and diseases are constrained from group overlapping. Newcomer et al. [355] and Goldstein et al. [185] overcome these limitations by clustering patients instead of diseases, but patients may not be excluded from clusters and agglomerative methods still do not provide a clear strategy on appropriate distance metrics and stopping criterion [344]. By using a semantic similarity measure with highest performance on established benchmarks, a more standardized approach is developed in this study for determining an appropriate distance metric. The OPTICS clustering method does not force every patient to join a cluster and focus can be placed on closely related cases. The Sander et al. [436] algorithm for hierarchically clustering OPTICS reachability plots provides a consistent method to accurately interpret and form patient groups from dendrograms.

A total of 22 clusters were identified containing 1,030 patients. Since patient ICU stays were mapped to SNOMED CT, groups could be classified by the most prevalent SNOMED CT diagnoses and procedures. Additionally, since SNOMED CT is a hierarchy, closely related concepts could be grouped into common parent concepts to give a clearer representation of the most frequent conditions in each cluster. As opposed to previous studies, similarities between patients were not based on a limited set of previously chosen high level chronic conditions. Each patient could potentially be identified by any of the 96,168 diagnoses in SNOMED CT. Additionally, common characteristics of patients can be determined from their procedures as well as their diagnoses.

Through further validation of clustering methods of patients with multimorbidities, data mining techniques will be able to more confidently discover nonrandom co-occurring conditions and the relationships behind their etiology. Future studies can take advantage of the additional wealth of information provided in EHRs by automatically mapping laboratory measurements, medications, procedures, and other clinical records to standard ontologies, such as LOINC [327], RxNorm [303], and CPT [25]. Natural language processing can be used to analyze and recognize terms in the unstructured data more deeply by incorporating the detailed notes in discharge summaries, radiology reports, pathology reports, and other narratives. The current study was able to identify multimorbidity clusters by only integration to SNOMED CT with patient diagnoses and procedures. There are many present initiatives, such as BioVU [419] and eMERGE [188] that are currently integrating patient records with genetic samples for analyses. With linking genetics to medical records, in-depth morbidity studies will be able to determine whether co-occurring diseases are related due to a common pathway from a single genetic variant [269]. As integration strategies of EHRs and related public resources continue to advance, more novel associations will be discovered to unravel the origins behind multimorbidity conditions.

CHAPTER VIII

CONCLUSION

In this dissertation, various methods were proposed towards the integration of biological and clinical data for the purposes of improving healthcare and discovering novel associations to disease. Most of the current biomedical knowledge is stored in distributed and heterogeneous data sources, and as a result, problems arise with data connectivity, integrity, redundancy, consistency and expandability [45]. Data warehousing is an ideal choice for researchers who wish to focus in a specific area by providing a quick and effective way to answer targeted queries. Greater control of curation and filtering features have the potential to deliver essential data quality. However, data volumes are usually too large to house locally, and it can be very complicated to create a global schema that accurately reflects all data types. Mediator based approaches always provide the most recent data without having to develop a global model, but there is not as much user control as in data warehousing. The availability of a mediated system is dependent on the availability of each external data source, and there can be large latencies in the retrieval of data. In Chapter 3, the BIOSPIDA system is introduced that alleviates many of the shortcomings of data warehousing. A universal parser can extract and transform any external data source as long as an XML or ASN.1 schema file is provided. No revisions are required to the universal parser if the data model changes for any of the data sources, and parallelization of the conversion speeds the creation of a fully integrated repository. Investments in hardware are less demanding with efficient memory management and redundancy removal eliminates large volumes of data to fractions of their original size. In a field that is constantly evolving, it is essential to develop a data integration solution that can keep up to date with external data model changes while delivering sufficient processing speeds.

Chapter 4 discusses an application towards the analysis of integrated biological data and healthcare by attempting to determine linkages between genes and diseases. The BioSPIDA system developed in Chapter 3 was utilized to create networks of gene annotations, including protein interactions, literature, expression, domains, pathways, and ontologies. With an integrated warehouse of diverse data types, associations to thousands of Homo Sapien genes could be quickly retrieved in seconds, facilitating the quick development of data for analysis. According to Rifaieh et al., scientists may spend 80% of their time retrieving and assembling their data [414]. The usage of data warehouse allowed queries to easily pre-process and deliver the needed information. All six networks were analyzed, and the diseases that were most predictive for each data type were reported.

The following chapters discuss the advantages of electronic health records (EHRs) and their enormous potential to transform healthcare into a data-rich domain. However, this potential is dependent on the ability to successfully combine primary data with external data sources at both the individual and population level. Many studies are poorly designed and do not properly utilize external datasets. The “Big Data” generated from medical practice is rarely captured and organized into an effective modeling platform. Biomedical terminologies were developed as a method to accurately represent knowledge about clinical events and healthcare data. Sophisticated analysis tools can be developed by recognizing structured phenotypes from heterogeneous patient data and mapping terms from clinical text to controlled medical vocabularies. Chapter 6 demonstrates an application of prioritizing ICU patients by severity and increasing the vocabulary of terms by integrating with the SNOMED CT ontology. This facilitated the creation of a severity model by conducting natural language processing on free text entries of diagnosis records. Chapters 5 and 7 illustrate a deeper integration of clinical data with public ontologies by creating a local data warehouse of MIMIC-II, UMLS and SNOMED CT. UMLS maps diagnosis and procedures entries to SNOMED CT and natural language processing is used to recognize unknown entries. By linking clinical records to standardized terminologies, the semantic distance between patients can be computed and similarity related clusters can be created. This helps to identify distinct multimorbidity groups and furthers the investigation of the functional relationships behind co-occurring conditions.

Each of these studies would not have been possible without the support of data integration technology, and it was found that the most suitable approach is dependent on the item of research. Semantic databases provide a machine-interpretable format where it is feasible to develop automatic reasoning methods. Natural language processing can recognize standardized concepts with an acceptable accuracy, but it is preferable to utilize mapped terminologies from structured and codified data when provided. Graph Databases can powerfully calculate sophisticated semantic distance between concepts when terms are linked to a publicly maintained ontology. The techniques of data warehousing, natural language processing, and graph databases have all contributed to creating a flexible data integration system for the biological and clinical systems. However, the next step in medical research is for the development of systems that can integrate both omics data and patient records in the creation of personalized healthcare.

Linking large-scale DNA databanks directly to electronic medical records could potentially transform healthcare by developing new rich genotype-phenotype associations [218, 339, 439]. These associations can be incorporated into patient care by reducing medication errors, improving diagnostic tests, and practice guidelines [167, 254,370,517] . A number of recent initiatives (eMERGE [188], BioVU [419]) have undertaken Electronic Health Driven Genome Research (EDGR) by recruiting patient cohorts and coupling medical records with microarray genotyping data. Previous studies have proven that successful results can be attained from EDGR by discovering novel genetic variants linked to disease [118], determining genomic regions associated with patient phenotypes [295], and detecting drug-gene interactions [122]. Despite these advancements, EDGR is faced with a number of challenges before it can be embraced more largely by healthcare centers. EHR-derived phenotypes coupled with DNA samples can provide sufficient enough level of information to discover and reproduce existing genetic variant disease associations. However, it is critical to accurately create definitions for cases and controls for each phenotype. Physicians frequently overturn incorrect diagnosis coding, and the sole usage of codes is not as reliable of a method for prediction of patient phenotypes. Frequently, natural language processing is required for information extraction of the unstructured data of clinical notes, inpatient histories, and discharge summaries. For example, concept recognition techniques may be needed to identify which medications, laboratory tests or diagnosis terms are necessary to determine a particular phenotype. The successful discovery of novel genotype-phenotype associations are also challenged by overlapping diseases, dataset sizes and selection biases. Overlapping diseases must be specified and excluded for each condition of interest [416]. Much larger sample sizes than those ordinarily attained may be required to accurately discern linkages between genotypes and phenotypes [230]. Datasets are not a selection from the whole world populations, and selection bias is a real possibility for all clinical studies [519].

Although great promise can emerge from the integration of genomic data with patient care, it simultaneously increases the complexity and amount of clinical data. Clinical phenotyping of patient records represent a “Big Data” problem in healthcare analytics, due to its intrinsic heterogeneity and enormous depth [97]. A popular definition of big data is proposed by Laney with the “3V” model [281]. Big data is described by three fundamental features of high volumes of data mass, high velocity of data flow, and high variety of data types. Big data in healthcare is especially challenging not only as a result of large data volumes, but also from the diversity of data types and the speed at which it must be managed [162]. However, efficient analysis and interpretation of big medical data can lead to new avenues for exploration and a better understanding of diseases with improved diagnostic and therapeutic strategies [141]. Potential benefits include detecting diseases at earlier stages and successful predictions of medical complications, lengths of stay, and elective surgery [406].

Big data analytics is typically broken down with distributed processing and executed across multiple nodes. Many big data tools have emerged from open-source development and Apache Hadoop[1] is one of the most significant distributed data processing platforms. The architecture of Hadoop originated from proprietary technology developed by Google to accommodate and index immense volumes of information, known as MapReduce [117]. Based on the Google File System [178], Hadoop can execute fault-tolerant distributed processing of large data sets across clusters of computers. This is facilitated with its implementation of the MapReduce algorithm and its own distributed database architecture, HBase[2], which was derived from the Google File System [488]. As the technology behind Hadoop continues to mature, new related software extend its applicability in the bioinformatics domain [406]. Mahout[3] is a machine learning library and can be used for the classification and clustering of data. However, the large scale and high dimensionality of patient records still provide infrastructure challenges for data mining, storage and integration. Cloud computing is a feasible solution where processing and storage capacity are dynamically provisioned. The cloud is a virtual machine (VM) prepackaged with needed software and executes tasks on parallel processors [83]. Many cloud computing services have been

[1] http://hadoop.apache.org/

[2] http://hbase.apache.org/

[3] https://mahout.apache.org/

emerging in the commercial sector including Google Compute Engine[4] and Amazon Elastic Compute Cloud (EC2)[5]. Numerous vendors provide open-source Hadoop platforms, including Amazon Web Services (AWS)[6], Cloudera[7], Hortonworks[8] and MapR[9], and many are available as cloud versions [359]. Cloud computing can accelerate the potential benefits of EHR sharing with the integration of genetic profiles and population-based information. By the storage of biomedical data on cloud servers without large hardware investments, resources and operation costs can be reduced limiting electronic health record startup expenses [83, 146, 209, 245, 448].

We currently live in an exciting period with many potential avenues for research in the health sciences that would not have been possible before the development of data integration, natural language processing, and "Big Data" technologies. All of these studies in this dissertation present potential areas where data analytics can improve health care. Future implementations will investigate the further advancements that can be made by linking even larger datasets of patient records to genomic regions. The progression of natural language processing will continue to improve its accuracy in recognizing clinical concepts from the unstructured notes of discharge summaries and patient histories. As more patient records are consolidated from many hospital visits, longitudinal studies will be available to develop a fuller understanding of the progression of disease. To process these large volumes of information, the arrival of "Big Data" technology and cloud computing could not come at a more perfect time for the clinical sciences.

[4] https://cloud.google.com/products/compute-engine/

[5] http://aws.amazon.com/ec2/

[6] http://aws.amazon.com/elasticmapreduce/

[7] http://www.cloudera.com/

[8] http://hortonworks.com/

[9] http://www.mapr.com/

REFERENCES

[1] Achard, F., Vaysseix, G., and Barillot, E., "Xml, bioinformatics and data integration," *Bioinformatics*, vol. 17, no. 2, pp. 115–125, 2001.

[2] Adamusiak, T., Shimoyama, N., and Shimoyama, M., "Next generation phenotyping using the unified medical language system," *JMIR Medical Informatics*, vol. 2, no. 1, p. e5, 2014.

[3] Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S., "Suspects: enabling fast and effective prioritization of positional candidates," *Bioinformatics*, vol. 22, no. 6, pp. 773–774, 2006.

[4] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., and others, "Gene prioritization through genomic data fusion," *Nature biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[5] Agorastos, T., Koutkias, V., Falelakis, M., Lekka, I., Mikos, T., Delopoulos, A., Mitkas, P. A., Tantsis, A., Weyers, S., Coorevits, P., and others, "Semantic integration of cervical cancer data repositories to facilitate multicenter association studies: the assist approach," *Cancer informatics*, vol. 8, p. 31, 2009.

[6] Al-Mubaid, H. and Nguyen, H. A., "A cluster-based approach for semantic similarity in the biomedical domain," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 2713–2717, IEEE, 2006.

[7] Amato, A. and Venticinque, S., "Big data management systems for the exploitation of pervasive environments," in *Big Data and Internet of Things: A Roadmap for Smart Environments*, pp. 67–89, Springer, 2014.

[8] Ambite, J. L., Ashish, N., Barish, G., Knoblock, C. A., Minton, S., Modi, P. J., Muslea, I., Philpot, A., and Tejada, S., "Ariadne: a system for constructing mediators for internet sources," in *ACM SIGMOD Record*, vol. 27, pp. 561–563, ACM, 1998.

[9] Angles, R. and Gutierrez, C., "Survey of graph database models," *ACM Computing Surveys (CSUR)*, vol. 40, no. 1, p. 1, 2008.

[10] Anguita, A., Martin, L., Perez-Rey, D., and Maojo, V., "A review of methods and tools for database integration in biomedicine," *Current Bioinformatics*, vol. 5, no. 4, pp. 253–269, 2010.

[11] Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J., "Optics: Ordering points to identify the clustering structure," in *ACM SIGMOD Record*, vol. 28, pp. 49–60, ACM, 1999.

[12] Antezana, E., Blondé, W., Egaña, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V., and Kuiper, M., "Biogateway: a semantic systems biology tool for the life sciences," *BMC bioinformatics*, vol. 10, no. Suppl 10, p. S11, 2009.

[13] Antezana, E., Egaña, M., Blondé, W., Illarramendi, A., Bilbao, I., De Baets, B., Stevens, R., Mironov, V., Kuiper, M., and others, "The cell cycle ontology: an application ontology for the representation and integrated analysis of the cell cycle process," *Genome Biol*, vol. 10, no. 5, p. R58, 2009.

[14] Antezana, E., Kuiper, M., and Mironov, V., "Biological knowledge management: the emerging role of the semantic web technologies," *Briefings in bioinformatics*, vol. 10, no. 4, pp. 392–407, 2009.

[15] Arenas, M., Kantere, V., Kementsietsidis, A., Kiringa, I., Miller, R. J., and Mylopoulos, J., "The hyperion project: from data integration to data coordination," *ACM SIGMOD Record*, vol. 32, no. 3, pp. 53–58, 2003.

[16] Aronson, A. R., "Effective mapping of biomedical text to the umls metathesaurus: the metamap program.," in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001.

[17] Aronson, A. R., "Metamap: Mapping text to the umls metathesaurus," *Bethesda, MD: NLM, NIH, DHHS*, 2006.

[18] Aronson, A. R. and Lang, F.-M., "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[19] Aseervatham, S. and Bennani, Y., "Semi-structured document categorization with a semantic kernel," *Pattern Recognition*, vol. 42, no. 9, pp. 2067–2076, 2009.

[20] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., and others, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[21] Aslam, J. F., "M.(2003) an information theoretic measure for document similarity," in *Proceedings of the 26the Annual International ACM SINGIR Conference on Research and Development in Information Retrieval (ACM Press) 449*, vol. 450.

[22] Association, A. D., *CDT: Current Dental Terminology.* American Dental Association, 2004.

[23] Association, A. H., *Hospital Statistics: 2000.* Health Forum, 1999.

[24] Association, A. H. and others, "Emergency department overload: A growing crisis. results of the aha survey of emergency department (ed) and hospital capacity," 2002.

[25] Association, A. M., *Current procedural terminology: CPT.* American Medical Association, 2007.

[26] Averill, R. F., Mullin, R. L., Steinbeck, B. A., Goldfield, N. I., and Grant, T. M., "Development of the icd-10 procedure coding system (icd-10-pcs).," *Topics in health information management*, vol. 21, no. 3, pp. 54–88, 2001.

[27] Bader, G. D., Betel, D., and Hogue, C. W., "Bind: the biomolecular interaction network database," *Nucleic acids research*, vol. 31, no. 1, pp. 248–250, 2003.

[28] Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., and others, "The universal protein resource (uniprot)," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D154–D159, 2005.

[29] Baker, P. G., Brass, A., Bechhofer, S., Goble, C. A., Paton, N. W., Stevens, R., and others, "Tambis: Transparent access to multiple bioinformatics information sources.," in *Ismb*, vol. 6, pp. 25–34, 1998.

[30] Ball, C. A., Sherlock, G., and Brazma, A., "Funding high-throughput data sharing," *Nature biotechnology*, vol. 22, no. 9, pp. 1179–1183, 2004.

[31] Barmpis, K. and Kolovos, D. S., "Comparative analysis of data persistence technologies for large-scale models," in *Proceedings of the 2012 Extreme Modeling Workshop*, pp. 33–38, ACM, 2012.

[32] Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., and Guthrie, B., "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study," *The Lancet*, vol. 380, no. 9836, pp. 37–43, 2012.

[33] Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R., "Ncbi geo: mining tens of millions of expression profilesdatabase and tools update," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D760–D765, 2007.

[34] Basu, M. K., Carmel, L., Rogozin, I. B., and Koonin, E. V., "Evolution of protein domain promiscuity in eukaryotes," *Genome research*, vol. 18, no. 3, pp. 449–461, 2008.

[35] Basu, M. K., Poliakov, E., and Rogozin, I. B., "Domain mobility in proteins: functional and evolutionary implications," *Briefings in bioinformatics*, vol. 10, no. 3, pp. 205–216, 2009.

[36] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., and others, "The pfam protein families database," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D138–D141, 2004.

[37] Batet, M., Sánchez, D., and Valls, A., "An ontology-based measure to compute semantic similarity in biomedicine," *Journal of biomedical informatics*, vol. 44, no. 1, pp. 118–125, 2011.

[38] Batra, S. and Tyagi, C., "Comparative analysis of relational and graph databases," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 2, pp. 509–512, 2012.

[39] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J., "Bio2rdf: towards a mashup to build bioinformatics knowledge systems," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 706–716, 2008.

[40] Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., and Sayers, E. W., "GenBank," *Nucleic Acids Res.*, vol. 40, pp. 48–53, Jan 2012.

[41] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L., "Genbank," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D25–D30, 2008.

[42] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E., "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[43] Berners-Lee, T., Hendler, J., Lassila, O., and others, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[44] Bird, S., "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72, Association for Computational Linguistics, 2006.

[45] Birkland, A. and Yona, G., "Biozon: a hub of heterogeneous biological data," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D235–D242, 2006.

[46] Biron, P., Malhotra, A., Consortium, W. W. W., and others, "Xml schema part 2: Datatypes," *World Wide Web Consortium Recommendation REC-xmlschema-2-20041028*, 2004.

[47] Björklund, Å. K., Ekman, D., Light, S., Frey-Skött, J., and Elofsson, A., "Domain rearrangements in protein evolution," *Journal of molecular biology*, vol. 353, no. 4, pp. 911–923, 2005.

[48] Bodenreider, O., "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[49] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., and others, "The swiss-prot protein knowledgebase and its supplement trembl in 2003," *Nucleic acids research*, vol. 31, no. 1, pp. 365–370, 2003.

[50] Bonafede, M. M., Suaya, J. A., Wilson, K. L., Mannino, D. M., and Polsky, D., "Incidence and cost of cap in a large working-age population.," *The American journal of managed care*, vol. 18, no. 7, pp. 380–387, 2012.

[51] Bonetta, L., "Getting up close and personal with your genome," *Cell*, vol. 133, no. 5, pp. 753–756, 2008.

[52] Botstein, D. and Risch, N., "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature genetics*, vol. 33, pp. 228–237, 2003.

[53] Boult, C., Dowd, B., McCaffrey, D., Boult, L., Hernandez, R., and Krulewitch, H., "Screening elders for risk of hospital admission.," *Journal of the American Geriatrics Society*, vol. 41, no. 8, p. 811, 1993.

[54] Box, D., Ehnebuske, D., Kakivaya, G., Layman, A., Mendelsohn, N., Nielsen, H. F., Thatte, S., and Winer, D., "Simple object access protocol (soap) 1.1," 2000.

[55] Boyd, C. M., Darer, J., Boult, C., Fried, L. P., Boult, L., and Wu, A. W., "Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases," *JAMA: The Journal of the American Medical Association*, vol. 294, no. 6, pp. 716–724, 2005.

[56] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F., "Extensible markup language (xml)," *World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210*, 1998.

[57] Brecher, C. and Spiezio, S., "Privatization and public hospitals: Choosing wisely for new york city," *New York Twentieth Century Fund Press*, vol. 103, p. 103, 1995.

[58] Bromberg, Y., "Disease gene prioritization," *PLoS computational biology*, vol. 9, no. 4, p. e1002902, 2013.

[59] Cabana, M. D., Rand, C. S., Powe, N. R., Wu, A. W., Wilson, M. H., Abboud, P.-A. C., and Rubin, H. R., "Why don't physicians follow clinical practice guidelines? a framework for improvement," *Jama*, vol. 282, no. 15, pp. 1458–1465, 1999.

[60] Calì, A., Calvanese, D., De Giacomo, G., and Lenzerini, M., "On the expressive power of data integration systems," in *Conceptual ModelingER 2002*, pp. 338–350, Springer, 2003.

[61] Canuel, V., Rance, B., Avillach, P., Degoulet, P., and Burgun, A., "Translational research platforms integrating clinical and omics data: a review of publicly available solutions," *Briefings in bioinformatics*, p. bbu006, 2014.

[62] Cao, H., Melton, G. B., Markatou, M., and Hripcsak, G., "Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases," *Journal of biomedical informatics*, vol. 41, no. 6, pp. 882–888, 2008.

[63] Caragea, D., Bao, J., Pathak, J., Silvescu, A., Andorf, C., Dobbs, D., and Honavar, V., "Information integration from semantically heterogeneous biological data sources," in *Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on*, pp. 580–584, IEEE, 2005.

[64] Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., Krummenacker, M., and others, "The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases," *Nucleic acids research*, vol. 38, no. suppl 1, pp. D473–D479, 2010.

[65] Centers for Medicare & Medicaid Services and others, *Healthcare Common Procedure Coding System (HCPCS)*. Centers for Medicare & Medicaid Services, 2003.

[66] Centers for Medicare and Medicaid Services, "Icd-10-cm official guidelines for coding and reporting. 2010," 2011.

[67] Centers for Medicare and Medicaid Services and others, "Icd-10-pcs reference manual," 2012.

[68] Cerami, E., *XML for Bioinformatics*. Springer, 2005.

[69] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., and others, "The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer discovery*, vol. 2, no. 5, pp. 401–404, 2012.

[70] Chakrabarty, R., Tran, T., Wolf, W., and Bickel, J., "Biostor: Developing an institutional biobank linked to the clinical record via i2b2," *AMIA Summit on Clinical Research Informatics*, vol. 2013, p. 1, 2013.

[71] Chalfin, D. B., Trzeciak, S., Likourezos, A., Baumann, B. M., and Dellinger, R. P., "Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit.," *Critical Care Medicine*, vol. 35, no. 6, pp. 1477–1483, 2007.

[72] Chan, A., Arendts, G., and Wong, S., "Causes of constraints to patient flow in emergency departments: a comparison between staff perceptions and findings from the patient flow study.," *Emergency medicine Australasia EMA*, vol. 20, no. 3, pp. 234–240, 2008.

[73] Chan, C. W., Bambos, N., and Escobar, G. J., "Maximizing throughput of hospital intensive care units with patient readmissions," *Working Paper, Columbia Business School*, pp. 1–41, 2011.

[74] Chan, L. W.-C., Liu, Y., Shyu, C.-R., and Benzie, I. F., "A snomed supported ontological vector model for subclinical disorder detection using ehr similarity," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 8, pp. 1398–1409, 2011.

[75] Chapman, B. and Chang, J., "Biopython: Python tools for computational biology," *ACM SIGBIO Newsletter*, vol. 20, no. 2, pp. 15–19, 2000.

[76] Chatr-Aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G., "Mint: the molecular interaction database," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D572–D574, 2007.

[77] Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M., and Plewczynski, D., "Ppi_svm: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables," *Cellular & molecular biology letters*, vol. 16, no. 2, pp. 264–278, 2011.

[78] Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., and Widom, J., "The tsimmis project: Integration of heterogenous information sources," 1994.

[79] Chen, C. and Hsu, C., "Indexing icd&#45;9 codes for free&#45;textual clinical diagnosis records by a new ensemble classifier," *Int. J. Comput. Intell. Bioinformatics Syst. Biol.*, vol. 1, pp. 177–192, December 2009.

[80] Chen, H., Mao, Y., Zheng, X., Cui, M., Feng, Y., Deng, S., Yin, A., Zhou, C., Tang, J., Jiang, X., and others, "Towards semantic e-science for traditional chinese medicine," *BMC bioinformatics*, vol. 8, no. Suppl 3, p. S6, 2007.

[81] Chen, H., Yu, T., and Chen, J. Y., "Semantic web meets integrative biology: a survey," *Briefings in bioinformatics*, vol. 14, no. 1, pp. 109–125, 2013.

[82] Chen, I., Min, A., and Markowitz, V. M., "An overview of the object protocol model (opm) and the opm data management tools," *Information Systems*, vol. 20, no. 5, pp. 393–418, 1995.

[83] Chen, J., Qian, F., Yan, W., and Shen, B., "Translational biomedical informatics in the cloud: present and future," *BioMed research international*, vol. 2013, 2013.

[84] Chen, J., Aronow, B. J., and Jegga, A. G., "Disease candidate gene identification and prioritization using protein interaction networks," *BMC bioinformatics*, vol. 10, no. 1, p. 73, 2009.

[85] Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G., "Toppgene suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic acids research*, vol. 37, no. suppl 2, pp. W305–W311, 2009.

[86] Chen, J., Xu, H., Aronow, B. J., and Jegga, A. G., "Improved human disease candidate gene prioritization using mouse phenotype," *BMC bioinformatics*, vol. 8, no. 1, p. 392, 2007.

[87] Chen, S. F. and Goodman, J., "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[88] Chen, Y., Wang, W., Zhou, Y., Shields, R., Chanda, S. K., Elston, R. C., and Li, J., "In silico gene prioritization by integrating multiple data sources," *PloS one*, vol. 6, no. 6, p. e21137, 2011.

[89] Chen, Y., Wu, X., and Jiang, R., "Integrating human omics data to prioritize candidate genes," *BMC medical genomics*, vol. 6, no. 1, p. 57, 2013.

[90] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., and others, "Sgd: Saccharomyces genome database," *Nucleic acids research*, vol. 26, no. 1, pp. 73–79, 1998.

[91] Cheung, K.-H., Frost, H. R., Marshall, M. S., Prud'hommeaux, E., Samwald, M., Zhao, J., and Paschke, A., "A journey to semantic web query federation in the life sciences," *BMC bioinformatics*, vol. 10, no. Suppl 10, p. S10, 2009.

[92] Cheung, K.-H., Yip, K. Y., Smith, A., Masiar, A., Gerstein, M., and others, "Yeasthub: a semantic web use case for integrating data in the life sciences domain," *Bioinformatics*, vol. 21, no. suppl 1, pp. i85–i96, 2005.

[93] Cheung, K.-H., Yip, K. Y., Townsend, J. P., and Scotch, M., "Hcls 2.0/3.0: Health care and life sciences data mashup using web 2.0/3.0," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 694–705, 2008.

[94] Chothia, C., Gough, J., Vogel, C., and Teichmann, S. A., "Evolution of the protein repertoire," *Science*, vol. 300, no. 5626, pp. 1701–1703, 2003.

[95] Chowbina, S., Hammamieh, R., Kumar, R., Chakraborty, N., Yang, R., Mudunuri, U., Jett, M., Palma, J. M., and Stephens, R., "Sysbiocube: A data warehouse and integrative data analysis platform facilitating systems biology studies of disorders of military relevance," *AMIA Summits on Translational Science Proceedings*, vol. 2013, p. 34, 2013.

[96] Chung, S. Y. and Wong, L., "Kleisli: a new tool for data integration in biology," *Trends in Biotechnology*, vol. 17, no. 9, pp. 351–355, 1999.

[97] Chute, C. G., Ullman-Cullere, M., Wood, G. M., Lin, S. M., He, M., and Pathak, J., "Some experiences and opportunities for big data in translational research," *Genetics in Medicine*, vol. 15, no. 10, pp. 802–809, 2013.

[98] Cimino, J. J. and Ayres, E. J., "The clinical research data repository of the us national institutes of health," *Studies in health technology and informatics*, vol. 160, no. Pt 2, p. 1299, 2010.

[99] Clancy, C. M., "Let the data be our guide: trends and tools for research on health care utilization," *Health economics*, vol. 21, no. 1, pp. 19–23, 2012.

[100] Cochran, J. K. and Bharti, A., "A multi-stage stochastic methodology for whole hospital bed planning under peak loading," *International Journal of Industrial and Systems Engineering*, vol. 1, no. 1/2, p. 8, 2006.

[101] Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and others, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[102] Codd, E. F., "A relational model of data for large shared data banks," *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, 1970.

[103] Cooke, C. R. and Iwashyna, T. J., "Using existing data to address important clinical questions in critical care," *Critical care medicine*, vol. 41, no. 3, pp. 886–896, 2013.

[104] Copley, R. R., Ponting, C. P., Schultz, J., and Bork, P., "Sequence analysis of multidomain proteins: past perspectives and future directions," *Advances in protein chemistry*, vol. 61, pp. 75–98, 2002.

[105] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C., "Section 22.4: Topological sort," in *Introduction to Algorithms*, pp. 549–552, MIT Press and McGraw-Hill, 2nd ed., 2001.

[106] Cornell, J. E., Pugh, J. A., Williams Jr, J. W., Kazis, L., Lee, A. F., Parchman, M. L., Zeber, J., Pederson, T., Montgomery, K. A., Noël, P. H., and others, "Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database," *Applied Multivariate Research*, vol. 12, no. 3, pp. 163–182, 2009.

[107] Couto, F. M. and Pinto, H. S., "The next generation of similarity measures that fully explore the semantics in biomedical ontologies," *Journal of bioinformatics and computational biology*, vol. 11, no. 05, 2013.

[108] Couto, F. M., Silva, M. J., and Coutinho, P. M., "Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 343–344, ACM, 2005.

[109] Couto, F. M., Silva, M. J., and others, "Disjunctive shared information between ontology concepts: application to gene ontology.," *J. Biomedical Semantics*, vol. 2, p. 5, 2011.

[110] Critchlow, T., Fidelis, K., Ganesh, M., Musick, R., and Slezak, T., "Datafoundry: information management for scientific data," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 4, no. 1, pp. 52–57, 2000.

[111] Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., and others, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 42, no. D1, pp. D472–D477, 2014.

[112] Daniel, C., Albuisson, E., Dart, T., Avillach, P., Cuggia, M., and Guo, Y., "Translational bioinformatics and clinical research informatics," in *Medical Informatics, e-Health*, pp. 429–461, Springer, 2014.

[113] Davidson, S. B., Crabtree, J., Brunk, B. P., Schug, J., Tannen, V., Overton, G. C., and Stoeckert Jr, C. J., "K2/kleisli and gus: Experiments in integrated access to genomic data sources," *IBM systems journal*, vol. 40, no. 2, pp. 512–531, 2001.

[114] Davidson, S. B., Overton, C., and Buneman, P., "Challenges in integrating biological data sources," *Journal of Computational Biology*, vol. 2, no. 4, pp. 557–572, 1995.

[115] Davidson, S. B., Overton, C., Tannen, V., and Wong, L., "Biokleisli: A digital library for biomedical researchers," *International Journal on Digital Libraries*, vol. 1, no. 1, pp. 36–53, 1997.

[116] De Bruijn, L. M., Hasman, A., and Arends, J. W., "Automatic snomed classification–a corpus-based method.," *Computer Methods and Programs in Biomedicine*, vol. 54, no. 1-2, pp. 115–122, 1997.

[117] Dean, J. and Ghemawat, S., "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[118] Denny, J. C., Crawford, D. C., Ritchie, M. D., Bielinski, S. J., Basford, M. A., Bradford, Y., Chai, H. S., Bastarache, L., Zuvich, R., Peissig, P., and others, "Variants near foxe1 are associated with hypothyroidism and other thyroid conditions: Using electronic medical records for genome-and phenome-wide studies," *The American Journal of Human Genetics*, vol. 89, no. 4, pp. 529–542, 2011.

[119] Denny, J. C., Miller, R. A., Johnson, K. B., and Spickard III, A., "Development and evaluation of a clinical note section header terminology," in *AMIA Annual Symposium proceedings*, vol. 2008, p. 156, American Medical Informatics Association, 2008.

[120] Denny, J. C., Spickard III, A., Johnson, K. B., Peterson, N. B., Peterson, J. F., and Miller, R. A., "Evaluation of a method to identify and categorize section headers in clinical documents," *Journal of the American Medical Informatics Association*, vol. 16, no. 6, pp. 806–815, 2009.

[121] Désarménien, J., "How to run TEX in french," Tech. Rep. SATN-CS-1013, Computer Science Department, Stanford University, Stanford, California, Aug. 1984.

[122] Dezentjé, V. O., van Blijderveen, N. J., Gelderblom, H., Putter, H., van Herk-Sukel, M. P., Casparie, M. K., Egberts, A. C., Nortier, J. W., and Guchelaar, H.-J., "Effect of concomitant cyp2d6 inhibitor use and tamoxifen adherence on breast cancer recurrence in early-stage breast cancer," *Journal of Clinical Oncology*, vol. 28, no. 14, pp. 2423–2429, 2010.

[123] Doan, S., Conway, M., Phuong, T. M., and Ohno-Machado, L., "Natural language processing in biomedicine: A unified system architecture overview," *arXiv preprint arXiv:1401.0569*, 2014.

[124] Dobson, G., Lee, H.-H., and Pinker, E., "A model of icu bumping," *Operations Research*, vol. 58, no. 6, pp. 1564–1576, 2010.

[125] Dominguez-Sal, D., Urbón-Bayes, P., Giménez-Vañó, A., Gómez-Villamor, S., Martínez-Bazán, N., and Larriba-Pey, J.-L., "Survey

of graph database performance on the hpc scalable graph analysis benchmark," in *Web-Age Information Management*, pp. 37–48, Springer, 2010.

[126] Donelson, L., Tarczy-Hornoch, P., Mork, P., Dolan, C., Mitchell, J. A., Barrier, M., and Mei, H., "The biomediator system as a data integration tool to answer diverse biologic queries," *Medinfo*, vol. 11, no. Pt 2, pp. 768–72, 2004.

[127] Dorr, D. A., Jones, S. S., Burns, L., Donnelly, S. M., Brunker, C. P., Wilcox, A., and Clayton, P. D., "Use of health-related, quality-of-life metrics to predict mortality and hospitalizations in community-dwelling seniors," *Journal of the American Geriatrics Society*, vol. 54, no. 4, pp. 667–673, 2006.

[128] Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R., and Stein, L., "The distributed annotation system," *BMC bioinformatics*, vol. 2, no. 1, p. 7, 2001.

[129] Durbin, C. G. and Kopel, R. F., "A case-control study of patients readmitted to the intensive care unit.," *Critical Care Medicine*, vol. 21, no. 10, pp. 1547–1553, 1993.

[130] Durbin, R. and Thierry-Mieg, J., "The acedb genome database," in *Computational methods in genome research*, pp. 45–55, Springer, 1994.

[131] Eckman, B. A., Kosky, A. S., and Laroco Jr, L. A., "Extending traditional query-based integration approaches for functional characterization of post-genomic data," *Bioinformatics*, vol. 17, no. 7, pp. 587–601, 2001.

[132] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D., "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.

[133] Elkin, P. L. and Tuttle, M. S., "Introduction," in *Terminology and Terminological Systems*, pp. 1–4, Springer, 2012.

[134] Elkin, P. L., Tuttle, M. S., Rallins, M., Trajkovski, J., Lumakovska, E., and Brown, S. H., "Implementations of terminology," in *Terminology and Terminological Systems*, pp. 125–175, Springer, 2012.

[135] Er, M., "A parallel computation approach to topological sorting," *The Computer Journal*, vol. 26, no. 4, pp. 293–295, 1983.

[136] Escobar, G. J. G. J., Greene, J. D. J. D., Scheirer, P., Gardner, M. N. M. N., Draper, D., and Kipnis, P., "Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases.," *Medical Care*, vol. 46, no. 3, pp. 232–9, 2008.

[137] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *KDD*, vol. 96, pp. 226–231, 1996.

[138] Etzold, T. and Argos, P., "Srsan indexing and retrieval tool for flat file data libraries," *Computer applications in the biosciences: CABIOS*, vol. 9, no. 1, pp. 49–57, 1993.

[139] Evans, R. S., Lloyd, J. F., and Pierce, L. A., "Clinical use of an enterprise data warehouse," in *AMIA Annual Symposium Proceedings*, vol. 2012, p. 189, American Medical Informatics Association, 2012.

[140] Everitt, B. S., Landau, S., Leese, M., and Stahl, D., "Cluster analysis: Wiley series in probability and statistics," 2011.

[141] Executive Office of the President, "Big data: Seizing Opportunities, Preserving Values." `http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf`. Accessed: 2014-07-12.

[142] Fallside, D. C. and Walmsley, P., "Xml schema part 0: primer second edition," *W3C recommendation*, p. 16, 2004.

[143] Fasshauer, G. E., "Positive definite kernels: past, present and future," *Dolomite Research Notes on Approximation*, vol. 4, pp. 21–63, 2011.

[144] Feldman, B., Martin, E. M., and Skotnes, T., "Big data in healthcare hype and hope," 2012.

[145] Fensel, D., van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Della Valle, E., Fischer, F., Huang, Z., Kiryakov, A., Lee, T.-I., and others, "Towards larkc: a platform for web-scale reasoning," in *Semantic Computing, 2008 IEEE International Conference on*, pp. 524–529, IEEE, 2008.

[146] Fernández-Cardeñosa, G., de la Torre-Díez, I., López-Coronado, M., and Rodrigues, J. J., "Analysis of cloud-based solutions on ehrs systems in different scenarios," *Journal of medical systems*, vol. 36, no. 6, pp. 3777–3782, 2012.

[147] Fernández-Suárez, X. M., Rigden, D. J., and Galperin, M. Y., "The 2014 nucleic acids research database issue and an updated nar online molecular biology database collection," *Nucleic acids research*, vol. 42, no. D1, pp. D1–D6, 2014.

[148] Fielding, R. T., *Architectural styles and the design of network-based software architectures*. PhD thesis, University of California, 2000.

[149] Fiks, A. G. and Grundmeier, R. W., "Elucidating challenges and opportunities in the transition to icd-10-cm," *Pediatrics*, pp. peds–2014, 2014.

[150] Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., and others, “Ensembl 2013,” *Nucleic acids research*, p. gks1236, 2012.

[151] Fong, J. H., Geer, L. Y., Panchenko, A. R., and Bryant, S. H., “Modeling the evolution of protein domain architectures using maximum parsimony,” *Journal of molecular biology*, vol. 366, no. 1, pp. 307–315, 2007.

[152] Formiga, F., Ferrer, A., Sanz, H., Marengoni, A., Alburquerque, J., and Pujol, R., “Patterns of comorbidity and multimorbidity in the oldest old: the octabaix study,” *European journal of internal medicine*, vol. 24, no. 1, pp. 40–44, 2013.

[153] Fortin, M., Bravo, G., Hudon, C., Lapointe, L., Almirall, J., Dubois, M.-F., and Vanasse, A., “Relationship between multimorbidity and health-related quality of life of patients in primary care,” *Quality of Life Research*, vol. 15, no. 1, pp. 83–91, 2006.

[154] Fortin, M., Soubhi, H., Hudon, C., Bayliss, E. A., and van den Akker, M., “Multimorbidity’s many challenges,” *BMJ: British Medical Journal*, vol. 334, no. 7602, p. 1016, 2007.

[155] Foster, I., *Designing and building parallel programs: concepts and tools for parallel software engineering*. Parallel programming / scientific computing, Addison-Wesley, 1995.

[156] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and others, “String v9. 1: protein-protein interaction networks, with increased coverage and integration,” *Nucleic acids research*, vol. 41, no. D1, pp. D808–D815, 2013.

[157] Franke, L., Bakel, H. v., Fokkens, L., De Jong, E. D., Egmont-Petersen, M., and Wijmenga, C., “Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes,” *The American Journal of Human Genetics*, vol. 78, no. 6, pp. 1011–1025, 2006.

[158] Freier, A., Hofestädt, R., Lange, M., Scholz, U., and Stephanik, A., “Biodataserver: a sql-based service for the online integration of life science data,” *In silico biology*, vol. 2, no. 2, pp. 37–57, 2002.

[159] Freudenberg, J. and Propping, P., “A similarity-based method for genome-wide prediction of disease-relevant human genes,” *Bioinformatics*, vol. 18, no. suppl 2, pp. S110–S115, 2002.

[160] Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., and Johnson, S. B., “A general natural-language text processor for clinical radiology,” *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 161–174, 1994.

[161] Friedman, M., Levy, A. Y., Millstein, T. D., and others, "Navigational plans for data integration," *AAAI/IAAI*, vol. 1999, pp. 67–73, 1999.

[162] Frost & Sullivan, "Drowning in Big Data? Reducing information technology complexities and costs for healthcare organizations.." `http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technology-complexities-ar.pdf`. Accessed: 2014-07-12.

[163] Fuchs, D., "The format of TEX's DVI files version 1," *TUGboat*, vol. 2, pp. 12–16, July 1981.

[164] Fuchs, D., "Device independent file format," *TUGboat*, vol. 3, pp. 14–19, Oct. 1982.

[165] Furney, S. J., Calvo, B., Larrañaga, P., Lozano, J. A., and Lopez-Bigas, N., "Prioritization of candidate cancer genesan aid to oncogenomic studies," *Nucleic acids research*, vol. 36, no. 18, pp. e115–e115, 2008.

[166] Furuta, R. K. and MacKay, P. A., "Two TEX implementations for the IBM PC," *Dr. Dobb's Journal*, vol. 10, pp. 80–91, Sept. 1985.

[167] Galanter, W. L., Hier, D. B., Jao, C., and Sarne, D., "Computerized physician order entry of medications and clinical decision support can improve problem list documentation compliance," *International journal of medical informatics*, vol. 79, no. 5, pp. 332–338, 2010.

[168] Galperin, M. Y., "The Molecular Biology Database Collection: 2007 update," *Nucleic Acids Res.*, vol. 35, pp. 3–4, Jan 2007.

[169] Gan, M., Dou, X., and Jiang, R., "From ontology to semantic similarity: calculation of ontology-based semantic similarity," *The Scientific World Journal*, vol. 2013, 2013.

[170] Gandhi, T., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., and others, "Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets," *Nature genetics*, vol. 38, no. 3, pp. 285–293, 2006.

[171] García-Olmos, L., Salvador, C. H., Alberquilla, Á., Lora, D., Carmona, M., García-Sagredo, P., Pascual, M., Muñoz, A., Monteagudo, J. L., and García-López, F., "Comorbidity patterns in patients with chronic diseases in general practice," *PloS one*, vol. 7, no. 2, p. e32141, 2012.

[172] Garin, N., Olaya, B., Perales, J., Moneta, M. V., Miret, M., Ayuso-Mateos, J. L., and Haro, J. M., "Multimorbidity patterns in a national

representative sample of the spanish adult population," *PloS one*, vol. 9, no. 1, p. e84794, 2014.

[173] Garla, V. N. and Brandt, C., "Semantic similarity in the biomedical domain: an evaluation across knowledge sources," *BMC bioinformatics*, vol. 13, no. 1, p. 261, 2012.

[174] Gaulton, K. J., Mohlke, K. L., and Vision, T. J., "A computational system to select candidate genes for complex human traits," *Bioinformatics*, vol. 23, no. 9, pp. 1132–1140, 2007.

[175] Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., Liu, C., Shi, W., and Bryant, S. H., "The ncbi biosystems database," *Nucleic acids research*, p. gkp858, 2009.

[176] Gentleman, R., "Visualizing and distances using go,"

[177] George, R. A., Liu, J. Y., Feng, L. L., Bryson-Richardson, R. J., Fatkin, D., and Wouters, M. A., "Analysis of protein sequence and interaction data for candidate disease gene prediction," *Nucleic acids research*, vol. 34, no. 19, pp. e130–e130, 2006.

[178] Ghemawat, S., Gobioff, H., and Leung, S.-T., "The google file system," in *ACM SIGOPS Operating Systems Review*, vol. 37, pp. 29–43, ACM, 2003.

[179] Giallourakis, C., Henson, C., Reich, M., Xie, X., and Mootha, V. K., "Disease gene discovery through integrative genomics," *Annu. Rev. Genomics Hum. Genet.*, vol. 6, pp. 381–406, 2005.

[180] Gijsen, R., Hoeymans, N., Schellevis, F. G., Ruwaard, D., Satariano, W. A., and van den Bos, G. A., "Causes and consequences of comorbidity: a review," *Journal of clinical epidemiology*, vol. 54, no. 7, pp. 661–674, 2001.

[181] Glazier, A. M., Nadeau, J. H., and Aitman, T. J., "Finding genes that underlie complex traits," *Science*, vol. 298, no. 5602, pp. 2345–2349, 2002.

[182] Goble, C. and Stevens, R., "State of the nation in data integration for bioinformatics," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 687–693, 2008.

[183] Goetghebeur, M. M., Forrest, S., and Hay, J. W., "Understanding the underlying drivers of inpatient cost growth: a literature review.," *The American Journal Of Managed Care*, vol. 9 Spec No 1, pp. SP3–P12, 2003.

[184] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L., "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[185] Goldstein, G., Luther, J. F., Jacoby, A. M., Haas, G. L., and Gordon, A. J., "A taxonomy of medical comorbidity for veterans who are homeless," *Journal of Health Care for the Poor and Underserved*, vol. 19, no. 3, pp. 991–1005, 2008.

[186] Golub, G. H. and Van Loan, C. F., *Matrix computations*, vol. 3. JHU Press, 2012.

[187] González, A. J. and Liao, L., "Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines," *BMC bioinformatics*, vol. 11, no. 1, p. 537, 2010.

[188] Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., Sanderson, S. C., Kannry, J., Zinberg, R., Basford, M. A., and others, "The electronic medical records and genomics (emerge) network: past, present, and future," *Genetics in Medicine*, vol. 15, no. 10, pp. 761–771, 2013.

[189] Graham, P. L. and Cook, D. A., "Prediction of risk of death using 30-day outcome: a practical end point for quality auditing in intensive care.," *Chest*, vol. 125, no. 4, pp. 1458–1466, 2004.

[190] Green, L., "Queueing analysis in healthcare," *Patient flow: reducing delay in healthcare delivery*, pp. 281–307, 2006.

[191] Green, L., "How many hospital beds?," *Journal Information*, vol. 39, no. 4, 2002.

[192] Green, L., "Using queueing theory to alleviate emergency department overcrowding," *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

[193] Gribble, S., Halevy, A., Ives, Z., Rodrig, M., and Suciu, D., "What can databases do for peer-to-peer,"

[194] Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M., *Fundamentals of queueing theory.* John Wiley & Sons, 2013.

[195] Gruber, T., "What is an ontology," 1993.

[196] Gupta, A., Ludascher, B., and Martone, M. E., "Knowledge-based integration of neuroscience data sources," in *Scientific and Statistical Database Management, 2000. Proceedings. 12th International Conference on*, pp. 39–52, IEEE, 2000.

[197] Guzzi, P. H., Mina, M., Guerra, C., and Cannataro, M., "Semantic similarity analysis of protein data: assessment with biological features and issues," *Briefings in bioinformatics*, vol. 13, no. 5, pp. 569–585, 2012.

[198] Haas, L., Kossmann, D., Wimmers, E., and Yang, J., "Optimizing queries across diverse data sources," 1997.

[199] Haas, L. M., Schwarz, P. M., Kodali, P., Kotlar, E., Rice, J. E., and Swope, W. C., "Discoverylink: A system for integrated access to life sciences data sources," *IBM systems Journal*, vol. 40, no. 2, pp. 489–511, 2001.

[200] Habib, S. L., Prihoda, T. J., Luna, M., and Werner, S. A., "Diabetes and risk of renal cell carcinoma," *Journal of Cancer*, vol. 3, p. 42, 2012.

[201] Hackett, M. L., Yapa, C., Parag, V., and Anderson, C. S., "Frequency of depression after stroke a systematic review of observational studies," *Stroke*, vol. 36, no. 6, pp. 1330–1340, 2005.

[202] Hagen, M. and Lee, E., "Biospida: A relational database translator for ncbi," in *AMIA Annual Symposium Proceedings*, vol. 2010, p. 422, American Medical Informatics Association, 2010.

[203] Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A., "Biomart central portalunified access to biological data," *Nucleic acids research*, vol. 37, no. suppl 2, pp. W23–W27, 2009.

[204] Halevy, A. Y., Ives, Z. G., Mork, P., and Tatarinov, I., "Piazza: Data management infrastructure for semantic web applications," in *Proceedings of the 12th international conference on World Wide Web*, pp. 556–567, ACM, 2003.

[205] Halland, K., Britz, K., and Gerber, A., "Investigations into the use of snomed ct to enhance an openmrs health information system," *South African Computer Journal*, vol. 47, pp. 33–45, 2011.

[206] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A., "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D514–D517, 2005.

[207] Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., and Montmain, J., "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain," *Journal of biomedical informatics*, 2013.

[208] Hart, K. W., Searls, D. B., and Overton, G. C., "Sortez: A relational translator for ncbi's asn. 1 database," *Computer applications in the biosciences: CABIOS*, vol. 10, no. 4, pp. 369–378, 1994.

[209] Haughton, J., "Look up: the right ehr may be in the cloud. major advantages include interoperability and flexibility.," *Health management technology*, vol. 32, no. 2, pp. 52–52, 2011.

[210] Hawkins, R. D., Hon, G. C., and Ren, B., "Next-generation genomics: an integrative approach," *Nature Reviews Genetics*, vol. 11, no. 7, pp. 476–486, 2010.

[211] Hebert, P. R., Gaziano, J. M., Chan, K. S., and Hennekens, C. H., "Cholesterol lowering with statin drugs, risk of stroke, and total mortality: an overview of randomized trials," *Jama*, vol. 278, no. 4, pp. 313–321, 1997.

[212] Henriksson, A., Conway, M., Duneld, M., and Chapman, W. W., "Identifying synonymy between snomed clinical terms of varying length using distributional analysis of electronic health records," in *AMIA Annual Symposium Proceedings*, vol. 2013, p. 600, American Medical Informatics Association, 2013.

[213] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., and others, "The hupo psi's molecular interaction formata community standard for the representation of protein interaction data," *Nature biotechnology*, vol. 22, no. 2, pp. 177–183, 2004.

[214] Hernandez, T. and Kambhampati, S., "Integration of biological sources: current systems and challenges ahead," *ACM SIgmod Record*, vol. 33, no. 3, pp. 51–60, 2004.

[215] Hersh, W., Jacko, J., Greenes, R., Tan, J., Janies, D., Embi, P., and Payne, P., "Health-care hit or miss?," *Nature*, vol. 470, no. 7334, pp. 327–329, 2011.

[216] HiMSS Analytics, "Applications of the EMRAM," *Essentials of the U.S. Hospital IT Market.* Winter 2013 8th Edition.

[217] Hliaoutakis, A., "Semantic similarity measures in mesh ontology and their application to information retrieval on medline," *Master's thesis*, 2005.

[218] Hoffman, M., "The genome-enabled electronic medical record," *Journal of Biomedical Informatics*, vol. 40, no. 1, pp. 44–46, 2007.

[219] Holden, L., Scuffham, P. A., Hilton, M. F., Muspratt, A., Ng, S.-K., and Whiteford, H. A., "Patterns of multimorbidity in working australians," *Population health metrics*, vol. 9, no. 1, p. 15, 2011.

[220] Hole, W. T., Carlsen, B. A., Tuttle, M. S., Srinivasan, S., Lipow, S. S., Olson, N. E., Sherertz, D. D., and Humphreys, B. L., "Achieving source transparency in the umls metathesaurus," *Medinfo*, vol. 11, no. Pt 1, pp. 371–5, 2004.

[221] Holford, M. E., McCusker, J. P., Cheung, K.-H., and Krauthammer, M., "A semantic web framework to integrate cancer omics data with biological knowledge," *BMC bioinformatics*, vol. 13, no. Suppl 1, p. S10, 2012.

[222] Hu, H., Correll, M., Kvecher, L., Osmond, M., Clark, J., Bekhash, A., Schwab, G., Gao, D., Gao, J., Kubatin, V., and others, "Dw4tr: a data warehouse for translational research," *Journal of biomedical informatics*, vol. 44, no. 6, pp. 1004–1019, 2011.

[223] Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., and Oinn, T., "Taverna: a tool for building and running workflows of services," *Nucleic acids research*, vol. 34, no. suppl 2, pp. W729–W732, 2006.

[224] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Pagni, M., and Sigrist, C. J., "The prosite database," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D227–D230, 2006.

[225] Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., and others, "Interpro: the integrative protein signature database," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D211–D215, 2009.

[226] Huttenhower, C., Hibbs, M. A., Myers, C. L., Caudy, A. A., Hess, D. C., and Troyanskaya, O. G., "The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction," *Bioinformatics*, vol. 25, no. 18, pp. 2404–2410, 2009.

[227] Improvement, N., "Managing multi-morbidity in practice what lessons can be learnt from the care of people with copd and co-morbidities?,"

[228] Inmon, B., "Data warehousing in a healthcare environment," *The Data Administration Newsletter-TDAN. com*, 2007.

[229] Inouye, S. K., Zhang, Y., Jones, R. N., Shi, P., Cupples, L. A., Calderon, H. N., and Marcantonio, E. R., "Risk factors for hospitalization among community-dwelling primary care older patients: development and validation of a predictive model.," *Medical care*, vol. 46, no. 7, pp. 726–731, 2008.

[230] Ioannidis, J. P., Trikalinos, T. A., and Khoury, M. J., "Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases," *American journal of epidemiology*, vol. 164, no. 7, pp. 609–614, 2006.

[231] Ishak, I. and Salim, N., "Database integration approaches for heterogeneous biological data sources: an overview," 2006.

[232] Jain, S. and Bader, G. D., "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC bioinformatics*, vol. 11, no. 1, p. 562, 2010.

[233] Jaiswal, N., *Priority queues.* Mathematics in science and engineering, Academic Press, 1968.

[234] Järvelin, K. and Kekäläinen, J., "Ir evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 41–48, ACM, 2000.

[235] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., and others, "String 8a global view on proteins and their functional interactions in 630 organisms," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D412–D416, 2009.

[236] Jensen, L. J., Lagarde, J., von Mering, C., and Bork, P., "Arrayprospector: a web resource of functional associations inferred from microarray expression data," *Nucleic acids research*, vol. 32, no. suppl 2, pp. W445–W448, 2004.

[237] Jensen, P. B., Jensen, L. J., and Brunak, S., "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[238] Jenssen, T.-K., Lægreid, A., Komorowski, J., and Hovig, E., "A literature network of human genes for high-throughput analysis of gene expression," *Nature genetics*, vol. 28, no. 1, pp. 21–28, 2001.

[239] Jiang, B.-B., Wang, J.-G., Wang, Y., and Xiao, J., "Gene prioritization for type 2 diabetes in tissue-specific protein interaction networks," *Syst Biol*, vol. 10801131, pp. 319–28, 2009.

[240] Jiang, J. J. and Conrath, D. W., "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[241] Jiang, R., Gan, M., and He, P., "Constructing a gene semantic similarity network for the inference of disease genes," *BMC systems biology*, vol. 5, no. Suppl 2, p. S2, 2011.

[242] Jimenez-Sanchez, G., Childs, B., and Valle, D., "Human disease genes," *Nature*, vol. 409, no. 6822, pp. 853–855, 2001.

[243] John, R., Kerby, D. S., and Hennessy, C. H., "Patterns and impact of comorbidity and multimorbidity among community-resident american indian elders," *The Gerontologist*, vol. 43, no. 5, pp. 649–660, 2003.

[244] Jones, A. R., Overly, C. C., and Sunkin, S. M., "The allen brain atlas: 5 years and beyond," *Nature Reviews Neuroscience*, vol. 10, no. 11, pp. 821–828, 2009.

[245] Kabachinski, J., "What's the forecast for cloud computing in healthcare?," *Biomedical Instrumentation & Technology*, vol. 45, no. 2, pp. 146–150, 2011.

[246] Kadam, U., "Redesigning the general practice consultation to improve care for patients with multimorbidity," *BMJ: British Medical Journal*, vol. 345, 2012.

[247] Kanehisa, M. and Bork, P., "Bioinformatics in the post-sequence era," *nature genetics*, vol. 33, pp. 305–310, 2003.

[248] Kanehisa, M. and Goto, S., "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[249] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., "The kegg resource for deciphering the genome," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D277–D280, 2004.

[250] Kann, M. G., "Advances in translational bioinformatics: computational approaches for the hunting of disease genes," *Briefings in bioinformatics*, vol. 11, no. 1, pp. 96–110, 2010.

[251] Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvych, N., Kurbatova, N., and others, "Gene expression atlas updatea value-added database of microarray and sequencing-based functional genomics experiments," *Nucleic acids research*, p. gkr913, 2011.

[252] Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., Rustici, G., Williams, E., Parkinson, H., and Brazma, A., "Gene expression atlas at the european bioinformatics institute," *Nucleic acids research*, p. gkp936, 2009.

[253] Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., and others, "The ucsc genome browser database," *Nucleic acids research*, vol. 31, no. 1, pp. 51–54, 2003.

[254] Kaushal, R., Jha, A. K., Franz, C., Glaser, J., Shetty, K. D., Jaggi, T., Middleton, B., Kuperman, G. J., Khorasani, R., Tanasijevic, M., and others, "Return on investment for a computerized physician order entry system," *Journal of the American Medical Informatics Association*, vol. 13, no. 3, pp. 261–266, 2006.

[255] Kaynar, A. M. and Sharma, S., "Respiratory failure," 2009.

[256] KC, D. and Terwiesch, C., "An econometric analysis of patient flows in the cardiac intensive care unit," *Manufacturing & Service Operations Management*, 2011.

[257] Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R., "Wikipathways: building research communities on biological pathways," *Nucleic acids research*, vol. 40, no. D1, pp. D1301–D1307, 2012.

[258] Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., and others, "Intactopen source resource for molecular interaction data," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D561–D565, 2007.

[259] Kim, J. Y. and Beckwith, B. A., "The coming wave of change: Icd-10," *Journal of pathology informatics*, vol. 1, 2010.

[260] Kirchberger, I., Meisinger, C., Heier, M., Zimmermann, A.-K., Thorand, B., Autenrieth, C. S., Peters, A., Ladwig, K.-H., and Döring, A., "Patterns of multimorbidity in the aged population. results from the kora-age study," *PloS one*, vol. 7, no. 1, p. e30556, 2012.

[261] Kirsten, T. and Rahm, E., "Biofuice: mapping-based data integration in bioinformatics," in *Data Integration in the Life Sciences*, pp. 124–135, Springer, 2006.

[262] Knuth, D. E., "The WEB system for structured documentation, version 2.3," Tech. Rep. STAN-CS-83-980, Computer Science Department, Stanford University, Stanford, California, Sept. 1983.

[263] Knuth, D. E., *The TEX Book*. Reading, Massachusetts: Addison-Wesley, 1984. Reprinted as Vol. A of *Computers & Typesetting*, 1986.

[264] Knuth, D. E., "Literate programming," *The Computer Journal*, vol. 27, pp. 97–111, May 1984.

[265] Knuth, D. E., "A torture test for TEX, version 1.3," Tech. Rep. STAN-CS-84-1027, Computer Science Department, Stanford University, Stanford, California, Nov. 1984.

[266] Knuth, D. E., *TEX: The Program*, vol. B of *Computers & Typesetting*. Reading, Massachusetts: Addison-Wesley, 1986.

[267] Koehler, S. B. and Adviser-Haug, P. J., *SymText: A natural language understanding system for encoding free text medical data*. The University of Utah, 1998.

[268] Koestler, T., von Haeseler, A., and Ebersberger, I., "Fact: Functional annotation transfer between proteins with similar feature architectures," *BMC bioinformatics*, vol. 11, no. 1, p. 417, 2010.

[269] Kohane, I. S., "Using electronic health records to drive discovery in disease genomics," *Nature Reviews Genetics*, vol. 12, no. 6, pp. 417–428, 2011.

[270] Köhler, J., Philippi, S., and Lange, M., "Semeda: ontology based semantic integration of biological databases," *Bioinformatics*, vol. 19, no. 18, pp. 2420–2427, 2003.

[271] Köhler, S., Bauer, S., Horn, D., and Robinson, P. N., "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[272] Kokangul, A., "A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit," *Computer methods and programs in biomedicine*, vol. 90, no. 1, pp. 56–65, 2008.

[273] Kolker, A., "Process modeling of icu patient flow: effect of daily load leveling of elective surgeries on icu diversion," *Journal of medical systems*, vol. 33, no. 1, pp. 27–40, 2009.

[274] Kondor, R. I. and Lafferty, J., "Diffusion kernels on graphs and other discrete input spaces,"

[275] Kubinger, K. D., "On artificial results due to using factor analysis for dichotomous variables," *Psychology Science*, vol. 45, no. 1, pp. 106–110, 2003.

[276] Kush, R. D., Helton, E., Rockhold, F. W., and Hardison, C. D., "Electronic health records, medical research, and the tower of babel," *New England Journal of Medicine*, vol. 358, no. 16, pp. 1738–1740, 2008.

[277] Lacroix, Z. and Critchlow, T., *Bioinformatics: managing scientific data*, vol. 6. Morgan Kaufmann, 2003.

[278] Lacroix, Z., Boucelma, O., and Essid, M., "The biological integration system," in *Proceedings of the 5th ACM international workshop on Web information and data management*, pp. 45–49, ACM, 2003.

[279] Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., and others, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[280] Lamport, L., *LaTeX: A Document Preparation System. User's Guide and Reference Manual.* Reading, Massachusetts: Addison-Wesley, 1986.

[281] Laney, D., "3d data management: Controlling data volume, velocity and variety," *META Group Research Note*, vol. 6, 2001.

[282] Leacock, C. and Chodorow, M., "Combining local context and wordnet similarity for word sense identification," *WordNet: An electronic lexical database*, vol. 49, no. 2, pp. 265–283, 1998.

[283] Lee, B. and Lee, D., "Dahunter: a web-based server that identifies homologous proteins by comparing domain architecture," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W60–W64, 2008.

[284] Lee, B. and Lee, D., "Protein comparison at the domain architecture level," *BMC bioinformatics*, vol. 10, no. Suppl 15, p. S5, 2009.

[285] Lee, D. and Chu, W. W., "Comparative analysis of six xml schema languages," *ACM Sigmod Record*, vol. 29, no. 3, pp. 76–87, 2000.

[286] Lee, E. K., Yuan, F., Hirsh, D. A., Mallory, M. D., and Simon, H. K., "A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department," in *AMIA Annual Symposium Proceedings*, vol. 2012, p. 495, American Medical Informatics Association, 2012.

[287] Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M., "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome research*, vol. 21, no. 7, pp. 1109–1121, 2011.

[288] Lee, T. J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D. W., Tenenbaum, J. D., and Karp, P. D., "Biowarehouse: a bioinformatics database warehouse toolkit," *BMC bioinformatics*, vol. 7, no. 1, p. 170, 2006.

[289] Lehmann, E. L. and D'Abrera, H. J., *Nonparametrics: statistical methods based on ranks.* Springer New York, 2006.

[290] Lemoine, F., Labedan, B., and Froidevaux, C., "Genoquery: a new querying module for functional annotation in a genomic warehouse," *Bioinformatics*, vol. 24, no. 13, pp. i322–i329, 2008.

[291] Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P., and Bork, P., "Recent improvements to the smart domain-based sequence annotation resource," *Nucleic acids research*, vol. 30, no. 1, pp. 242–244, 2002.

[292] Levy, A., Rajaraman, A., and Ordille, J., "Querying heterogeneous information sources using source descriptions," 1996.

[293] Li, Y. and Patra, J. C., "Integration of multiple data sources to prioritize candidate genes using discounted rating system," *BMC bioinformatics*, vol. 11, no. Suppl 1, p. S20, 2010.

[294] Li, Y., Bandar, Z. A., and McLean, D., "An approach for measuring semantic similarity between words using multiple information sources," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 871–882, 2003.

[295] Liao, K. P., Kurreeman, F., Li, G., Duclos, G., Murphy, S., Guzman, R., Cai, T., Gupta, N., Gainer, V., Schur, P., and others, "Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the

electronic medical records in rheumatoid arthritis cases and non–rheumatoid arthritis controls," *Arthritis & Rheumatism*, vol. 65, no. 3, pp. 571–581, 2013.

[296] Liekens, A. M., De Knijf, J., Daelemans, W., Goethals, B., De Rijk, P., Del-Favero, J., and others, "Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation," *Genome Biol*, vol. 12, no. 6, p. R57, 2011.

[297] Lin, D., "An information-theoretic definition of similarity.," in *ICML*, vol. 98, pp. 296–304, 1998.

[298] Lindberg, D. A., Humphreys, B. L., and McCray, A. T., "The unified medical language system.," *Methods of information in medicine*, vol. 32, no. 4, pp. 281–291, 1993.

[299] Linghu, B., Snitkin, E. S., Hu, Z., Xia, Y., and DeLisi, C., "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," *Genome Biol*, vol. 10, no. 9, p. R91, 2009.

[300] Litvak, E., "Optimizing patient flow by managing its variability," in *From front office to front line: essential issues for health care leaders*, pp. 91–111, Joint Commission Resources, 2005.

[301] Litvak, E. and Long, M., "Cost and quality under managed care: Irreconcilable differences," *Am J Manag Care*, vol. 6, no. 3, pp. 305–12, 2000.

[302] Liu, H., Wu, S. T., Li, D., Jonnalagadda, S., Sohn, S., Wagholikar, K., Haug, P. J., Huff, S. M., and Chute, C. G., "Towards a semantic lexicon for clinical natural language processing," in *AMIA Annual Symposium Proceedings*, vol. 2012, p. 568, American Medical Informatics Association, 2012.

[303] Liu, S., Ma, W., Moore, R., Ganesan, V., and Nelson, S., "Rxnorm: prescription for electronic drug information exchange," *IT professional*, vol. 7, no. 5, pp. 17–23, 2005.

[304] Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., and Tarczy-Hornoch, P., "Data integration and genomic medicine," *Journal of biomedical informatics*, vol. 40, no. 1, pp. 5–16, 2007.

[305] Lowe, H. J., Ferris, T. A., Hernandez, P. M., and Weber, S. C., "Stride–an integrated standards-based translational research informatics platform," in *AMIA Annual Symposium Proceedings*, vol. 2009, p. 391, American Medical Informatics Association, 2009.

[306] Luijks, H. D., Loeffen, M. J., Lagro-Janssen, A. L., van Weel, C., Lucassen, P. L., and Schermer, T. R., "Gps' considerations in multimorbidity management: a qualitative study," *British Journal of General Practice*, vol. 62, no. 600, pp. e503–e510, 2012.

[307] Ma, X., Lee, H., Wang, L., and Sun, F., "Cgi: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data," *Bioinformatics*, vol. 23, no. 2, pp. 215–221, 2007.

[308] Mabotuwana, T., Lee, M. C., and Cohen-Solal, E. V., "An ontology-based similarity measure for biomedical data–application to radiology reports," *Journal of biomedical informatics*, vol. 46, no. 5, pp. 857–868, 2013.

[309] Machado, C. M., Rebholz-Schuhmann, D., Freitas, A. T., and Couto, F. M., "The semantic web in translational medicine: current applications and future directions," *Briefings in bioinformatics*, p. bbt079, 2013.

[310] MacQueen, J. and others, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, p. 14, California, USA, 1967.

[311] Madhavan, S., Gusev, Y., Harris, M., Tanenbaum, D. M., Gauba, R., Bhuvaneshwar, K., Shinohara, A., Rosso, K., Carabet, L. A., Song, L., and others, "G-doc: a systems medicine platform for personalized oncology," *Neoplasia (New York, NY)*, vol. 13, no. 9, p. 771, 2011.

[312] Madylova, A. and Oguducu, S., "A taxonomy based semantic similarity of documents using the cosine measure," in *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, pp. 129–134, IEEE, 2009.

[313] Maglott, D., Ostell, J., Pruitt, K., and Tatusova, T., "Entrez gene: gene-centered information at ncbi," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D52–D57, 2011.

[314] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T., "Entrez gene: gene-centered information at ncbi," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D54–D58, 2005.

[315] Manconi, A. and Rodriguez-Tomé, P., "A survey on integrating data in bioinformatics," in *Learning Structure and Schemas from Documents*, pp. 413–432, Springer, 2011.

[316] Mangin, D., Heath, I., and Jamoulle, M., "Beyond diagnosis: rising to the multimorbidity challenge," *BMJ: British Medical Journal*, vol. 344, 2012.

[317] Marchler-Bauer, A., Lu, S., Anderson, J., Chitsaz, F., Derbyshire, M., DeWeese-Scott, C., Fong, J., Geer, L., Geer, R., Gonzales, N., and others, "Cdd: a conserved domain database for the functional annotation of proteins," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D225–D229, 2011.

[318] Marengoni, A., Bonometti, F., Nobili, A., Tettamanti, M., Salerno, F., Corrao, S., Iorio, A., Marcucci, M., and Mannucci, P., "In-hospital

death and adverse clinical events in elderly patients according to disease clustering: the reposi study," *Rejuvenation research*, vol. 13, no. 4, pp. 469–477, 2010.

[319] Marengoni, A., Rizzuto, D., Wang, H.-X., Winblad, B., and Fratiglioni, L., "Patterns of chronic multimorbidity in the elderly population," *Journal of the American Geriatrics Society*, vol. 57, no. 2, pp. 225–230, 2009.

[320] McBride, B., "The resource description framework (rdf) and its vocabulary description language rdfs," in *Handbook on Ontologies*, pp. 51–65, Springer, 2004.

[321] McCaig, L. F. and Burt, C. W., "National Hospital Ambulatory Medical Care Survey: 2001 emergency department summary," *Adv Data*, pp. 1–29, Jun 2003.

[322] McCaig, L. F. and Burt, C. W., "National Hospital Ambulatory Medical Care Survey: 2002 emergency department summary," *Adv Data*, pp. 1–34, Mar 2004.

[323] McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., and others, "The emerge network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies," *BMC medical genomics*, vol. 4, no. 1, p. 13, 2011.

[324] McConnell, P., Dash, R. C., Chilukuri, R., Pietrobon, R., Johnson, K., Annechiarico, R., and Cuticchia, A. J., "The cancer translational research informatics platform," *BMC medical informatics and decision making*, vol. 8, no. 1, p. 60, 2008.

[325] McCray, A. and Divita, G., "Asn. 1: defining a grammar for the umls knowledge sources.," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, p. 868, American Medical Informatics Association, 1995.

[326] McCusker, J. P., Phillips, J. A., Beltrán, A. G., Finkelstein, A., and Krauthammer, M., "Semantic web data warehousing for cagrid," *BMC bioinformatics*, vol. 10, no. Suppl 10, p. S2, 2009.

[327] McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., and others, "Loinc, a universal standard for identifying laboratory observations: a 5-year update," *Clinical chemistry*, vol. 49, no. 4, pp. 624–633, 2003.

[328] McGuinness, D. L., Van Harmelen, F., and others, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 2004-03, p. 10, 2004.

[329] McManus, M. L., Long, M. C., Cooper, A., and Litvak, E., "Queuing theory accurately models the need for critical care resources.," *Anesthesiology*, vol. 100, no. 5, pp. 1271–1276, 2004.

[330] McShea, M., Holl, R., Badawi, O., Riker, R., and Silfen, E., "The eicu research institute-a collaboration between industry, health-care providers, and academia," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 29, no. 2, pp. 18–25, 2010.

[331] Melton, G. B., Parsons, S., Morrison, F. P., Rothschild, A. S., Markatou, M., and Hripcsak, G., "Inter-patient distance metrics using snomed ct defining relationships," *Journal of biomedical informatics*, vol. 39, no. 6, pp. 697–705, 2006.

[332] Melton, G. B., Raman, N., Chen, E. S., Sarkar, I. N., Pakhomov, S., and Madoff, R. D., "Evaluation of family history information within clinical documents and adequacy of hl7 clinical statement and clinical genomics family history models for its representation: a case report," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 337–340, 2010.

[333] Mewes, H.-W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., and Weil, B., "Mips: a database for genomes and protein sequences," *Nucleic acids research*, vol. 30, no. 1, pp. 31–34, 2002.

[334] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., Hurdle, J. F., and others, "Extracting information from textual documents in the electronic health record: a review of recent research," *Yearb Med Inform*, vol. 35, pp. 128–44, 2008.

[335] Miled, Z. B., Li, N., Baumgartner, M., and Liu, Y., "A decentralized approach to the integration of life science web databases," *INFORMATICA-LJUBLJANA-*, vol. 27, no. 1, pp. 3–14, 2003.

[336] Miled, Z. B., Li, N., and Bukhres, O., "Baciis: biological and chemical information integration system," *Journal of Database Management (JDM)*, vol. 16, no. 3, pp. 72–85, 2005.

[337] Miller, E., "An introduction to the resource description framework," *Bulletin of the American Society for Information Science and Technology*, vol. 25, no. 1, pp. 15–19, 1998.

[338] Miller, G. A. and Charles, W. G., "Contextual correlates of semantic similarity," *Language and cognitive processes*, vol. 6, no. 1, pp. 1–28, 1991.

[339] Mitchell, D. R. and Mitchell, J. A., "Status of clinical gene sequencing data reporting and associated risks for information loss," *Journal of biomedical informatics*, vol. 40, no. 1, pp. 47–54, 2007.

[340] Miyoshi, N. S. B., Pinheiro, D. G., Silva, W. A., and Felipe, J. C., "Computational framework to support integration of biomolecular and clinical data within a translational approach," *BMC bioinformatics*, vol. 14, no. 1, p. 180, 2013.

[341] Mootha, V., Lepage, P., Miller, K., Bunkenborg, J., Reich, M., Hjerrild, M., Delmonte, T., Villeneuve, A., Sladek, R., Xu, F., and others, "Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics," *Proceedings of the National Academy of Sciences*, vol. 100, no. 2, p. 605, 2003.

[342] Mordelet, F. and Vert, J.-P., "Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples," *BMC bioinformatics*, vol. 12, no. 1, p. 389, 2011.

[343] Moreau, Y. and Tranchevent, L.-C., "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 523–536, 2012.

[344] Morgan, B. J. and Ray, A. P., "Non-uniqueness and inversions in cluster analysis," *Applied statistics*, pp. 117–134, 1995.

[345] Mork, P., Halevy, A., and Tarczy-Hornoch, P., "A model for data integration systems of biomedical data applied to online genetic databases.," in *Proceedings of the AMIA Symposium*, p. 473, American Medical Informatics Association, 2001.

[346] Mulaik, S., *Foundations of factor analysis*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series, CRC Press, 2009.

[347] Müller, H. and Freytag, J.-C., *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005.

[348] Muller, K. and Vignaux, T., "Simpy: Simulating systems in python," *ONLamp. com Python Devcenter*, 2003.

[349] Murphy, S. N., Mendis, M., Hackett, K., Kuttan, R., Pan, W., Phillips, L. C., Gainer, V., Berkowicz, D., Glaser, J. P., Kohane, I., and others, "Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside," in *AMIA annual symposium proceedings*, vol. 2007, p. 548, American Medical Informatics Association, 2007.

[350] Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., and Kohane, I., "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," *Journal of the American Medical Informatics Association*, vol. 17, no. 2, pp. 124–130, 2010.

[351] Murtagh, F., "A survey of recent advances in hierarchical clustering algorithms," *The Computer Journal*, vol. 26, no. 4, pp. 354–359, 1983.

[352] Musen, M., Wieckert, K., Miller, E., Campbell, K., and Fagan, L., "Development of a controlled medical terminology: knowledge acquisition and knowledge representation.," *Methods of information in medicine*, vol. 34, no. 1-2, pp. 85–95, 1995.

[353] National Center for Health Statistics (US) and others, *ICD-9-CM: International Classification of Diseases 9th Revision Clinical Modification.* US Department of Health and Human Services, Public Health Service, Health Care Financing Administration, 1980.

[354] Neogi, S., Krestyaninova, M., Kapushesky, M., Emam, I., Brazma, A., and others, "Moda-a data warehouse for multi-omics data," *J Data Mining Genomics Proteomics*, vol. 4, no. 145, pp. 2153–0602, 2013.

[355] Newcomer, S. R., Steiner, J. F., and Bayliss, E. A., "Identifying subgroups of complex patients with cluster analysis.," *The American journal of managed care*, vol. 17, no. 8, pp. e324–32, 2010.

[356] Ng, S.-K., Zhang, Z., Tan, S.-H., and Lin, K., "Interdom: a database of putative interacting protein domains for validating predicted protein interactions and complexes," *Nucleic acids research*, vol. 31, no. 1, pp. 251–254, 2003.

[357] Ng, S. K., Holden, L., and Sun, J., "Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics," *Statistics in medicine*, vol. 31, no. 27, pp. 3393–3405, 2012.

[358] Noy, N. F., "Semantic integration: a survey of ontology-based approaches," *ACM Sigmod Record*, vol. 33, no. 4, pp. 65–70, 2004.

[359] Ohlhorst, F. J., *Big data analytics: turning big data into big money.* John Wiley & Sons, 2012.

[360] Ohno-Machado, L., "Realizing the full potential of electronic health records: the role of natural language processing," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 539–539, 2011.

[361] Ohno-Machado, L., Bafna, V., Boxwala, A. A., Chapman, B. E., Chapman, W. W., Chaudhuri, K., Day, M. E., Farcas, C., Heintzman, N. D., Jiang, X., and others, "idash: integrating data for analysis, anonymization, and sharing," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 196–201, 2012.

[362] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A., and others,

"Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, no. 17, pp. 3045–3054, 2004.

[363] Olson, K., Hayduk, L., and Thomas, J., "Comparing two approaches for studying symptom clusters: factor analysis and structural equation modeling," *Supportive Care in Cancer*, vol. 22, no. 1, pp. 153–161, 2014.

[364] on A Framework for Developing a New Taxonomy of Disease, N. R. C. U. C. and others, *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease.* National Academies Press (US), 2011.

[365] Organization, W. H. and others, "International classification of impairments, disabilities, and handicaps: a manual of classification relating to the consequences of disease, published in accordance with resolution wha29. 35 of the twenty-ninth world health assembly, may 1976," 1980.

[366] Ortutay, C. and Vihinen, M., "Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies," *Nucleic acids research*, vol. 37, no. 2, pp. 622–628, 2009.

[367] Ostell, J., "Databases of discovery," *Queue*, vol. 3, no. 3, pp. 40–48, 2005.

[368] Ostell, J. M. and Kans, J. A., "The ncbi data model," *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Volume 39*, pp. 121–144, 1998.

[369] Ostell, J., "The entrez search and retrieval system," *The NCBI Handbook[Internet], National Library of Medicine (US), National Center for Biotechnology Information, Bethesda, MD*, 2002.

[370] Ozcan, Y. A. and Kazley, A. S., "Do hospitals with electronic medical records (emrs) provide higher quality care? an examination of three clinical conditions," *Medical Care Research and Review*, 2008.

[371] Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. B., "Semantic similarity and relatedness between clinical terms: an experimental study," in *AMIA annual symposium proceedings*, vol. 2010, p. 572, American Medical Informatics Association, 2010.

[372] Pakhomov, S., Buntrock, J., and Chute, C., "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," *Journal of the American Medical Informatics Association*, vol. 13, no. 5, pp. 516–525, 2006.

[373] Patashnik, O., *BibTEXing.* Computer Science Department, Stanford University, Stanford, California, Jan. 1988. Available in the BibTEX release.

[374] Patashnik, O., *Designing BibTEX Styles.* Computer Science Department, Stanford University, Jan. 1988.

[375] Patrick, J., Wang, Y., and Budd, P., "An automated system for conversion of clinical notes into snomed clinical terminology," in *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*, pp. 219–226, Australian Computer Society, Inc., 2007.

[376] Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G., "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of biomedical informatics*, vol. 40, no. 3, pp. 288–299, 2007.

[377] Pekar, V. and Staab, S., "Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1–7, Association for Computational Linguistics, 2002.

[378] Peng, H. and Rui, J., "Integrating multiple gene semantic similarity profiles to infer disease genes," in *Control Conference (CCC), 2012 31st Chinese*, pp. 7420–4725, IEEE, 2012.

[379] Pérez, J., Arenas, M., and Gutierrez, C., "Semantics and complexity of sparql," in *The Semantic Web-ISWC 2006*, pp. 30–43, Springer, 2006.

[380] Perez-Iratxeta, C., Bork, P., and Andrade, M. A., "Association of genes to genetically inherited diseases using data mining," *Nature genetics*, vol. 31, no. 3, pp. 316–319, 2002.

[381] Perez-Iratxeta, C., Wjst, M., Bork, P., and Andrade, M. A., "G2d: a tool for mining genes associated with disease," *BMC genetics*, vol. 6, no. 1, p. 45, 2005.

[382] Pérez-Rey, D., Maojo, V., García-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martín-Sánchez, F., and Sousa, A., "Ontofusion: Ontology-based integration of genomic and clinical databases," *Computers in Biology and Medicine*, vol. 36, no. 7, pp. 712–730, 2006.

[383] Pers, T. H., Hansen, N. T., Lage, K., Koefoed, P., Dworzynski, P., Miller, M. L., Flint, T. J., Mellerup, E., Dam, H., Andreassen, O. A., and others, "Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes," *Genetic epidemiology*, vol. 35, no. 5, pp. 318–332, 2011.

[384] Pesquita, C., Faria, D., Bastos, H., Falcão, A., and Couto, F., "Evaluating go-based semantic similarity measures," in *Proc. 10th Annual Bio-Ontologies Meeting*, pp. 37–40, 2007.

[385] Pesquita, C., Faria, D., Falcao, A. O., Lord, P., and Couto, F. M., "Semantic similarity in biomedical ontologies," *PLoS computational biology*, vol. 5, no. 7, p. e1000443, 2009.

[386] Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K., and Duch, W., "A shared task involving multi-label classification of clinical free text," in *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pp. 97–104, Association for Computational Linguistics, 2007.

[387] Petrakis, E. G., Varelas, G., Hliaoutakis, A., and Raftopoulou, P., "X-similarity: computing semantic similarity between concepts from different ontologies," *JDIM*, vol. 4, no. 4, pp. 233–237, 2006.

[388] Pico, A. R., Kelder, T., Van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C., "Wikipathways: pathway editing for the people," *PLoS biology*, vol. 6, no. 7, p. e184, 2008.

[389] Piro, R. M. and Di Cunto, F., "Computational approaches to disease-gene prediction: rationale, classification and successes," *FEBS Journal*, vol. 279, no. 5, pp. 678–696, 2012.

[390] Pirró, G. and Euzenat, J., "A feature and information theoretic framework for semantic similarity and relatedness," in *The Semantic Web–ISWC 2010*, pp. 615–630, Springer, 2010.

[391] Pivovarov, R. and Elhadad, N., "A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts," *Journal of biomedical informatics*, vol. 45, no. 3, pp. 471–481, 2012.

[392] Plaza, L. and Díaz, A., "Retrieval of similar electronic health records using umls concept graphs," in *Natural Language Processing and Information Systems*, pp. 296–303, Springer, 2010.

[393] Poblador-Plou, B., Calderón-Larrañaga, A., Marta-Moreno, J., Hancco-Saavedra, J., Sicras-Mainar, A., Soljak, M., and Prados-Torres, A., "Comorbidity of dementia: a cross-sectional study of primary care older patients," *BMC psychiatry*, vol. 14, no. 1, p. 84, 2014.

[394] Poland, R. and Ostrea, E., "Neonatal hyperbilirubinemia," *Care of the high risk neonate (3rd ed), Saunders, Philadelphia*, pp. 239–261, 1986.

[395] Prados-Torres, A., Calderón-Larrañaga, A., Hancco-Saavedra, J., Poblador-Plou, B., and van den Akker, M., "Multimorbidity patterns: a systematic review," *Journal of clinical epidemiology*, vol. 67, no. 3, pp. 254–266, 2014.

[396] Prados-Torres, A., Poblador-Plou, B., Calderón-Larrañaga, A., Gimeno-Feliu, L. A., González-Rubio, F., Poncel-Falcó, A., Sicras-Mainar, A., and Alcalá-Nalvaiz, J. T., "Multimorbidity patterns in primary care: interactions among chronic diseases using factor analysis," *PloS one*, vol. 7, no. 2, p. e32190, 2012.

[397] Prasad, T. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., and others, "Human protein reference database2009 update," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D767–D772, 2009.

[398] Pruitt, K., Tatusova, T., Klimke, W., and Maglott, D., "Ncbi reference sequences: current status, policy and new initiatives," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D32–D36, 2009.

[399] Pruitt, K. D. and Maglott, D. R., "Refseq and locuslink: Ncbi gene-centered resources," *Nucleic acids research*, vol. 29, no. 1, pp. 137–140, 2001.

[400] Pruitt, K. D., Tatusova, T., and Maglott, D. R., "Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D61–D65, 2007.

[401] Punta, M. and Ofran, Y., "The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function," *PLoS computational biology*, vol. 4, no. 10, p. e1000160, 2008.

[402] Quackenbush, J., "Computational analysis of microarray data," *Nature reviews genetics*, vol. 2, no. 6, pp. 418–427, 2001.

[403] Rada, R., Mili, H., Bicknell, E., and Blettner, M., "Development and application of a metric on semantic nets," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 19, no. 1, pp. 17–30, 1989.

[404] Radivojac, P., Peng, K., Clark, W. T., Peters, B. J., Mohan, A., Boyle, S. M., and Mooney, S. D., "An integrated approach to inferring gene–disease associations in humans," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, no. 3, pp. 1030–1037, 2008.

[405] Raghavachari, B., Tasneem, A., Przytycka, T. M., and Jothi, R., "Domine: a database of protein domain interactions," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D656–D661, 2008.

[406] Raghupathi, W. and Raghupathi, V., "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, p. 3, 2014.

[407] Rebholz-Schuhmann, D., Grabmüller, C., Kavaliauskas, S., Croset, S., Woollard, P., Backofen, R., Filsell, W., and Clark, D., "A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources," *Drug discovery today*, 2013.

[408] Redelmeier, D. A., Tan, S. H., and Booth, G. L., "The treatment of unrelated disorders in patients with chronic medical diseases," *New England Journal of Medicine*, vol. 338, no. 21, pp. 1516–1520, 1998.

[409] Reimand, J., Hui, S., Jain, S., Law, B., and Bader, G. D., "Domain-mediated protein interaction prediction: From genome to network," *FEBS letters*, vol. 586, no. 17, pp. 2751–2763, 2012.

[410] Render, M. L., Freyberg, R. W., Hasselbeck, R., Hofer, T. P., Sales, A. E., Deddens, J., Levesque, O., and Almenoff, P. L., "Infrastructure for quality transformation: measurement and reporting in veterans administration intensive care units," *BMJ quality & safety*, vol. 20, no. 6, pp. 498–507, 2011.

[411] Resnik, P., "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint cmp-lg/9511007*, 1995.

[412] Richardson, D., "The access-block effect: relationship between delay to reaching an inpatient bed and inpatient length of stay," *Medical Journal of Australia*, vol. 177, no. 9, pp. 492–495, 2002.

[413] Ridge, J., Jones, S., Nielsen, M., and Shahani, A., "Capacity planning for intensive care units," *European journal of operational research*, vol. 105, no. 2, pp. 346–355, 1998.

[414] Rifaieh, R., Unwin, R., Carver, J., and Miller, M., "Swami: integrating biological databases and analysis tools within user friendly environment," in *Proceedings of the 4th international conference on Data integration in the life sciences*, pp. 48–58, Springer-Verlag, 2007.

[415] Ripamonti, C., De Conno, F., Ventafridda, V., Rossi, B., and Baines, M., "Management of bowel obstruction in advanced and terminal cancer patients," *Annals of oncology*, vol. 4, no. 1, pp. 15–21, 1993.

[416] Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., Basford, M. A., Brown-Gentry, K., Balser, J. R., Masys, D. R., and others, "Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record," *The American Journal of Human Genetics*, vol. 86, no. 4, pp. 560–572, 2010.

[417] Ritter, O., "The integrated genomic database (igd)," in *Computational Methods in Genome Research*, pp. 57–73, Springer, 1994.

[418] Robinson, I., Webber, J., and Eifrem, E., *Graph Databases.* O'Reilly Media, 2013.

[419] Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., and Masys, D. R., "Development of a large-scale de-identified dna biobank to enable personalized medicine," *Clinical Pharmacology & Therapeutics*, vol. 84, no. 3, pp. 362–369, 2008.

[420] Rodríguez, M. A. and Egenhofer, M. J., "Determining semantic similarity among entity classes from different ontologies," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 2, pp. 442–456, 2003.

[421] Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., and others, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS computational biology*, vol. 7, no. 8, p. e1002141, 2011.

[422] Rosenbloom, S. T., Denny, J. C., Xu, H., Lorenzi, N., Stead, W. W., and Johnson, K. B., "Data from clinical notes: a perspective on the tension between structure and flexible documentation," *Journal of the American Medical Informatics Association*, vol. 18, no. 2, pp. 181–186, 2011.

[423] Rothberg, M. B., Pekow, P. S., Lahti, M., Brody, O., Skiest, D. J., and Lindenauer, P. K., "Antibiotic therapy and treatment failure in patients hospitalized for acute exacerbations of chronic obstructive pulmonary disease," *Jama*, vol. 303, no. 20, pp. 2035–2042, 2010.

[424] Ruch, P., Gobeill, J., Tbahriti, I., and Geissbühler, A., "From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding," in *AMIA Annual Symposium Proceedings*, vol. 2008, p. 636, American Medical Informatics Association, 2008.

[425] Ruttenberg, A., Rees, J. A., Samwald, M., and Marshall, M. S., "Life sciences on the semantic web: the neurocommons and beyond," *Briefings in bioinformatics*, vol. 10, no. 2, pp. 193–204, 2009.

[426] Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G., "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database," *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.

[427] Sager, N., Lyman, M., Bucknall, C., Nhan, N., and Tick, L. J., "Natural language processing and the representation of clinical data," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 142–160, 1994.

[428] Sahama, T. R. and Croll, P. R., "A data warehouse architecture for clinical data warehousing," in *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*, pp. 227–232, Australian Computer Society, Inc., 2007.

[429] Salton, G., Wong, A., and Yang, C.-S., "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[430] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D., "The database of interacting proteins: 2004 update," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D449–D451, 2004.

[431] Samuel, A. L., "First grade TEX: A beginner's TEX manual," Tech. Rep. SATN-CS-83-985, Computer Science Department, Stanford University, Stanford, California, Nov. 1983.

[432] SáNchez, D. and Batet, M., "Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective," *Journal of biomedical informatics*, vol. 44, no. 5, pp. 749–759, 2011.

[433] Sánchez, D., Batet, M., and Isern, D., "Ontology-based information content computation," *Knowledge-Based Systems*, vol. 24, no. 2, pp. 297–303, 2011.

[434] Sánchez, D., Batet, M., Isern, D., and Valls, A., "Ontology-based semantic similarity: A new feature-based approach," *Expert Systems with Applications*, vol. 39, no. 9, pp. 7718–7728, 2012.

[435] Sander, J., Ester, M., Kriegel, H.-P., and Xu, X., "Density-based clustering in spatial databases: The algorithm gdbscan and its applications," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 169–194, 1998.

[436] Sander, J., Qin, X., Lu, Z., Niu, N., and Kovarsky, A., "Automatic extraction of clusters from hierarchical clustering representations," in *Advances in Knowledge Discovery and Data Mining*, pp. 75–87, Springer, 2003.

[437] Šarić, J., Jensen, L. J., Ouzounova, R., Rojas, I., and Bork, P., "Extraction of regulatory gene/protein networks from medline," *Bioinformatics*, vol. 22, no. 6, pp. 645–650, 2006.

[438] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G., "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[439] Sax, U. and Schmidt, S., "Integration of genomic data in electronic health records-opportunities and dilemmas," *Methods of information in medicine*, vol. 44, no. 4, p. 546, 2005.

[440] Sayers, E., "The e-utilities in-depth: parameters, syntax and more," 2013.

[441] Sayers, E., "A general introduction to the e-utilities," 2013.

[442] Sayers, E., Barrett, T., Benson, D., Bolton, E., Bryant, S., Canese, K., Chetvernin, V., Church, D., DiCuccio, M., Federhen, S., and others, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D38–D51, 2011.

[443] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H., "Pid: the pathway interaction database," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D674–D679, 2009.

[444] Schäfer, I., von Leitner, E.-C., Schön, G., Koller, D., Hansen, H., Kolonko, T., Kaduszkiewicz, H., Wegscheider, K., Glaeske, G., and van den Bussche, H., "Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions," *PloS one*, vol. 5, no. 12, p. e15941, 2010.

[445] Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T., "A new measure for functional similarity of gene products based on gene ontology," *BMC bioinformatics*, vol. 7, no. 1, p. 302, 2006.

[446] Schlicker, A., Lengauer, T., and Albrecht, M., "Improving disease gene prioritization using the semantic similarity of gene ontology terms," *Bioinformatics*, vol. 26, no. 18, pp. i561–i567, 2010.

[447] Schölkopf, B., Tsuda, K., and Vert, J.-P., *Kernel methods in computational biology.* MIT press, 2004.

[448] Schweitzer, E. J., "Reconciliation of the cloud computing model with us federal electronic health record regulations," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 161–165, 2012.

[449] Scott, K. S., *Health Information Management Technology: An Applied Approach*, ch. 5. Clinical Vocabularies and Classification Systems. American Health Information Management Association, 2012.

[450] Seco, N., Veale, T., and Hayes, J., "An intrinsic information content metric for semantic similarity in wordnet," in *ECAI*, vol. 16, p. 1089, Citeseer, 2004.

[451] Seelow, D., Schwarz, J. M., and Schuelke, M., "Genedistillerdistilling candidate genes from linkage intervals," *PLoS One*, vol. 3, no. 12, p. e3874, 2008.

[452] Segagni, D., Tibollo, V., Dagliati, A., Perinati, L., Zambelli, A., Priori, S., and Bellazzi, R., "The onco-i2b2 project: integrating biobank information and clinical data to support translational research in oncology.," *Studies in health technology and informatics*, vol. 169, pp. 887–891, 2010.

[453] Shah, S., Huang, Y., Xu, T., Yuen, M., Ling, J., and Ouellette, B., "Atlas–a data warehouse for integrative bioinformatics," *BMC bioinformatics*, vol. 6, no. 1, p. 34, 2005.

[454] Shahani, A., Korve, N., Jones, K., and Paynton, D., "Towards an operational model for prevention and treatment of asthma attacks," *Journal of the Operational Research Society*, pp. 916–926, 1994.

[455] Shi, X., "Semantic web services: an unfulfilled promise,"

[456] Shimokawa, K., Mogushi, K., Shoji, S., Hiraishi, A., Ido, K., Mizushima, H., and Tanaka, H., "icod: an integrated clinical omics database based on the systems-pathology view of disease," *BMC genomics*, vol. 11, no. Suppl 4, p. S19, 2010.

[457] Shin, S.-Y., Kim, W. S., and Lee, J.-H., "Characteristics desired in clinical data warehouse for biomedical research," *Healthcare informatics research*, vol. 20, no. 2, pp. 109–116, 2014.

[458] Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., and Lai, A. M., "A review of approaches to identifying patient phenotype cohorts using electronic health records," *Journal of the American Medical Informatics Association*, pp. amiajnl–2013, 2013.

[459] Siddharthan, K., Jones, W., and Johnson, J., "A priority queuing model to reduce waiting times in emergency care," *International Journal of Health Care Quality Assurance*, vol. 9, no. 5, pp. 10–16, 1996.

[460] Siddiqi, J., Akhgar, B., Gruzdz, A., Zaefarian, G., and Ihnatowicz, A., "Automated diagnosis system to support colon cancer treatment: Match," in *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on*, pp. 201–205, IEEE, 2008.

[461] Simpson, M. S. and Demner-Fushman, D., "Biomedical text mining: A survey of recent progress," in *Mining Text Data*, pp. 465–517, Springer, 2012.

[462] Sinha, P., Sunder, G., Bendale, P., Mantri, M., and Dande, A., "Extended coding system for patient care procedures," *Electronic Health Record: Standards, Coding Systems, Frameworks, and Infrastructures*, pp. 131–135.

[463] Smith, N. G. and Eyre-Walker, A., "Human disease genes: patterns and predictions," *Gene*, vol. 318, pp. 169–175, 2003.

[464] Snel, B., Lehmann, G., Bork, P., and Huynen, M. A., "String: a webserver to retrieve and display the repeatedly occurring neighbourhood of a gene," *Nucleic acids research*, vol. 28, no. 18, pp. 3442–3444, 2000.

[465] Snow, N., Bergin, K., Horrigan, T., and others, "Readmission of patients to the surgical intensive care unit: patient profiles and possibilities for prevention.," *Critical care medicine*, vol. 13, no. 11, p. 961, 1985.

[466] Song, N., Joseph, J. M., Davis, G. B., and Durand, D., "Sequence similarity network reveals common ancestry of multidomain proteins," *PLoS computational biology*, vol. 4, no. 5, p. e1000063, 2008.

[467] Spivak, M. D., *The Joy of TeX*. American Mathematical Society, 1985.

[468] Sprivulis, P., Da Silva, J., Jacobs, I., Frazer, A., Jelinek, G., and others, "The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments," *Medical Journal of Australia*, vol. 184, no. 5, p. 208, 2006.

[469] Starfield, B., "New paradigms for quality in primary care.," *The British Journal of General Practice*, vol. 51, no. 465, p. 303, 2001.

[470] Starfield, B., Lemke, K. W., Bernhardt, T., Foldes, S. S., Forrest, C. B., and Weiner, J. P., "Comorbidity: implications for the importance of primary care in casemanagement," *The Annals of Family Medicine*, vol. 1, no. 1, pp. 8–14, 2003.

[471] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M., "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[472] Stearns, M. Q., Price, C., Spackman, K. A., and Wang, A. Y., "Snomed clinical terms: overview of the development process and project status.," in *Proceedings of the AMIA Symposium*, p. 662, American Medical Informatics Association, 2001.

[473] Stein, L. D., "Integrating biological databases," *Nature Reviews Genetics*, vol. 4, no. 5, pp. 337–345, 2003.

[474] Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., and others, "The generic genome browser: a building block for a model organism system database," *Genome research*, vol. 12, no. 10, pp. 1599–1610, 2002.

[475] Stevens, R. D., Robinson, A. J., and Goble, C. A., "mygrid: personalised bioinformatics on the information grid," *Bioinformatics*, vol. 19, no. suppl 1, pp. i302–i304, 2003.

[476] Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., and others, "Population genomics of human gene expression," *Nature genetics*, vol. 39, no. 10, pp. 1217–1224, 2007.

[477] Stuart, J. M., Segal, E., Koller, D., and Kim, S. K., "A gene-coexpression network for global discovery of conserved genetic modules," *science*, vol. 302, no. 5643, pp. 249–255, 2003.

[478] Sugumaran, V. and Storey, V. C., "Ontologies for conceptual modeling: their creation, use, and management," *Data & knowledge engineering*, vol. 42, no. 3, pp. 251–271, 2002.

[479] Sujansky, W., "Heterogeneous database integration in biomedicine," *Journal of biomedical informatics*, vol. 34, no. 4, pp. 285–298, 2001.

[480] Sun, J. and Zhao, Z., "A comparative study of cancer proteins in the human protein-protein interaction network," *BMC genomics*, vol. 11, no. Suppl 3, p. S5, 2010.

[481] Syamaladevi, D. P., Joshi, A., and Sowdhamini, R., "An alignment-free domain architecture similarity search (adass) algorithm for inferring homology between multi-domain proteins," *Bioinformation*, vol. 9, no. 10, p. 491, 2013.

[482] Szalma, S., Koka, V., Khasanova, T., and Perakslis, E. D., "Effective knowledge management in translational medicine," *Journal of translational medicine*, vol. 8, no. 1, p. 68, 2010.

[483] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., and others, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D561–D568, 2011.

[484] Tan, A., Tripp, B., and Daley, D., "Briskresearch-oriented storage kit for biology-related data," *Bioinformatics*, vol. 27, no. 17, pp. 2422–2425, 2011.

[485] Tan, P.-N., Steinbach, M., and Kumar, V., "Cluster analysis: basic concepts and algorithms," *Introduction to data mining*, pp. 487–568, 2006.

[486] Tatarinov, I., Ives, Z., Madhavan, J., Halevy, A., Suciu, D., Dalvi, N., Dong, X. L., Kadiyska, Y., Miklau, G., and Mork, P., "The piazza peer data management project," *ACM Sigmod Record*, vol. 32, no. 3, pp. 47–52, 2003.

[487] Tatonetti, N., Denny, J., Murphy, S., Fernald, G., Krishnan, G., Castro, V., Yue, P., Tsau, P., Kohane, I., Roden, D., and others, "Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels," *Clinical Pharmacology & Therapeutics*, vol. 90, no. 1, pp. 133–142, 2011.

[488] Taylor, R. C., "An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics," *BMC bioinformatics*, vol. 11, no. Suppl 12, p. S1, 2010.

[489] Terris, D. D., Litaker, D. G., and Koroukian, S. M., "Health state information derived from secondary databases is affected by multiple sources of bias," *Journal of clinical epidemiology*, vol. 60, no. 7, pp. 734–741, 2007.

[490] Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., and others, "Panther: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification," *Nucleic acids research*, vol. 31, no. 1, pp. 334–341, 2003.

[491] Thompson, H. S., Beech, D., Maloney, M., and Mendelsohn, N., "Xml schema part 1: structures second edition," 2004.

[492] Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., and Hide, W. A., "Integration of text-and data-mining using ontologies successfully selects disease gene candidates," *Nucleic acids research*, vol. 33, no. 5, pp. 1544–1552, 2005.

[493] Tinetti, M., Bogardus Jr, S., and Agostini, J., "Potential pitfalls of disease-specific guidelines for patients with multiple conditions.," *The New England journal of medicine*, vol. 351, no. 27, pp. 2870–2874, 2004.

[494] Töpel, T., Kormeier, B., Klassen, A., and Hofestädt, R., "Biodwh: a data warehouse kit for life science data integration," *Journal of integrative bioinformatics*, vol. 5, no. 2, p. 93, 2008.

[495] Tranchevent, L.-C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S., and Moreau, Y., "Endeavour update: a web resource for gene prioritization in multiple species," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W377–W384, 2008.

[496] Tu, J., Mazer, C., Levinton, C., Armstrong, P., and Naylor, C., "A predictive index for length of stay in the intensive care unit following cardiac surgery.," *CMAJ: Canadian Medical Association Journal*, vol. 151, no. 2, p. 177, 1994.

[497] Turner, F. S., Clutterbuck, D. R., Semple, C. A., and others, "Pocus: mining genomic sequence annotation to predict disease genes," *Genome biology*, vol. 4, no. 11, pp. R75–R75, 2003.

[498] Tversky, A., "Features of similarity.," *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[499] Ullman, J. D., "Information integration using logical views," in *Database TheoryICDT'97*, pp. 19–40, Springer, 1997.

[500] Uzuner, Ö., "Recognizing obesity and comorbidities in sparse data," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 561–570, 2009.

[501] Uzuner, Ö., Goldstein, I., Luo, Y., and Kohane, I., "Identifying patient smoking status from medical discharge records," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 14–24, 2008.

[502] Uzuner, Ö., Luo, Y., and Szolovits, P., "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, 2007.

[503] Uzuner, Ö., Solti, I., and Cadag, E., "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 514–518, 2010.

[504] Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L., "2010 i2b2/va challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, 2011.

[505] Vafaee, F., Rosu, D., Broackes-Carter, F., and Jurisica, I., "Novel semantic similarity measure improves an integrative approach to predicting gene functional associations," *BMC systems biology*, vol. 7, no. 1, p. 22, 2013.

[506] Vakatov, D., "Configure, build, and use the toolkit," 2013.

[507] van den Akker, M., Buntinx, F., and Knottnerus, J. A., "Comorbidity or multimorbidity: what's in a name? a review of literature," *European Journal of General Practice*, vol. 2, no. 2, pp. 65–70, 1996.

[508] van den Akker, M., Buntinx, F., Metsemakers, J. F., Roos, S., and Knottnerus, J. A., "Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases," *Journal of clinical epidemiology*, vol. 51, no. 5, pp. 367–375, 1998.

[509] van den Akker, M., Buntinx, F., Roos, S., and Knottnerus, J. A., "Problems in determining occurrence rates of multimorbidity," *Journal of clinical epidemiology*, vol. 54, no. 7, pp. 675–679, 2001.

[510] van den Bussche, H., Koller, D., Kolonko, T., Hansen, H., Wegscheider, K., Glaeske, G., von Leitner, E.-C., Schäfer, I., and Schön, G., "Which chronic diseases and disease combinations are specific to multimorbidity in the elderly? results of a claims data based cross-sectional study in germany," *BMC public health*, vol. 11, no. 1, p. 101, 2011.

[511] Vandervalk, B., McCarthy, L., and Wilkinson, M., "Cardioshare: Web services for the semantic web," *Semantic Web Challenge*, vol. 2008, 2008.

[512] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R., "Associating genes and protein complexes with disease via network propagation," *PLoS computational biology*, vol. 6, no. 1, p. e1000641, 2010.

[513] Verbrugge, L. M., Lepkowski, J. M., and Imanaka, Y., "Comorbidity and its impact on disability," *The Milbank Quarterly*, pp. 450–484, 1989.

[514] Von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B., "String: a database of predicted functional associations between proteins," *Nucleic acids research*, vol. 31, no. 1, pp. 258–261, 2003.

[515] Von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B., and Bork, P., "String 7recent developments in the integration and prediction of protein interactions," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D358–D362, 2007.

[516] Von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P., "String: known and predicted protein–protein associations, integrated and transferred across organisms," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D433–D437, 2005.

[517] Walsh, K. E., Landrigan, C. P., Adams, W. G., Vinci, R. J., Chessare, J. B., Cooper, M. R., Hebert, P. M., Schainker, E. G., McLaughlin, T. J., and Bauchner, H., "Effect of computer order entry on prevention of serious medication errors in hospitalized children," *Pediatrics*, vol. 121, no. 3, pp. e421–e427, 2008.

[518] Wang, L., Riethoven, J., and Robinson, A., "Xembl: distributing embl data in xml format," *Bioinformatics*, vol. 18, no. 8, pp. 1147–1148, 2002.

[519] Wang, W. and Krishnan, E., "Big data and clinicians: a review on the state of the science," *JMIR Medical Informatics*, vol. 2, no. 1, p. e1, 2014.

[520] Wang, Y., Addess, K., Chen, J., Geer, L., He, J., He, S., Lu, S., Madej, T., Marchler-Bauer, A., Thiessen, P., and others, "Mmdb: annotating protein sequences with entrez's 3d-structure database," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D298–D300, 2007.

[521] Wang, Y., Xiao, J., Suzek, T., Zhang, J., Wang, J., and Bryant, S., "Pubchem: a public information system for analyzing bioactivities of small molecules," *Nucleic acids research*, vol. 37, no. suppl 2, pp. W623–W633, 2009.

[522] Wang, Z., Gao, X., He, C., Miller, J., Kissinger, J., Heiges, M., Aurrecoechea, C., Kraemer, E., and Pennington, C., "A comparison of federated databases with web services for the integration of bioinformatics data," in *Proc. of the 2007 Int. Conference on Bioinformatics & Computational Biology (BIOCOMP), Las Vegas, NV (June 2007) pp*, pp. 334–338, Citeseer, 2007.

[523] Wiederhold, G., "Mediators in the architecture of future information systems," *Computer*, vol. 25, no. 3, pp. 38–49, 1992.

[524] Wilkinson, M. D. and Links, M., "Biomoby: an open source biological web services proposal," *Briefings in bioinformatics*, vol. 3, no. 4, pp. 331–341, 2002.

[525] Wilson, D., Madera, M., Vogel, C., Chothia, C., and Gough, J., "The superfamily database in 2007: families and functions," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D308–D313, 2007.

[526] Wilson, P. W., DAgostino, R. B., Parise, H., Sullivan, L., and Meigs, J. B., "Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus," *Circulation*, vol. 112, no. 20, pp. 3066–3072, 2005.

[527] Wolff, J. L., Starfield, B., and Anderson, G., "Prevalence, expenditures, and complications of multiple chronic conditions in the elderly," *Archives of internal medicine*, vol. 162, no. 20, pp. 2269–2276, 2002.

[528] Wong, A., Boshuizen, H. C., Schellevis, F. G., Kommer, G. J., and Polder, J. J., "Longitudinal administrative data can be used to examine multimorbidity, provided false discoveries are controlled for," *Journal of clinical epidemiology*, vol. 64, no. 10, pp. 1109–1117, 2011.

[529] World Health Organization, *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*, vol. 1. World Health Organization, 1992.

[530] Wu, X., Pang, E., Lin, K., and Pei, Z.-M., "Improving the measurement of semantic similarity between gene ontology terms and gene products: Insights from an edge-and ic-based hybrid method," *PloS one*, vol. 8, no. 5, p. e66745, 2013.

[531] Wu, X., Zhu, L., Guo, J., Zhang, D.-Y., and Lin, K., "Prediction of yeast protein–protein interaction network: insights from the gene ontology and annotations," *Nucleic acids research*, vol. 34, no. 7, pp. 2137–2150, 2006.

[532] Wu, X., Jiang, R., Zhang, M. Q., and Li, S., "Network-based global inference of human disease genes," *Molecular systems biology*, vol. 4, no. 1, 2008.

[533] Wu, Z. and Palmer, M., "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133–138, Association for Computational Linguistics, 1994.

[534] Wunsch, H., Guerra, C., Barnato, A. E., Angus, D. C., Li, G., and Linde-Zwirble, W. T., "Three-year outcomes for medicare beneficiaries who survive intensive care," *Jama*, vol. 303, no. 9, pp. 849–856, 2010.

[535] Wunsch, H., Harrison, D. A., and Rowan, K., "Health services research in critical care using administrative data," *Journal of critical care*, vol. 20, no. 3, pp. 264–269, 2005.

[536] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D., “Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions,” *Nucleic acids research*, vol. 30, no. 1, pp. 303–305, 2002.

[537] Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., and Denny, J. C., “Medex: a medication information extraction system for clinical narratives,” *Journal of the American Medical Informatics Association*, vol. 17, no. 1, pp. 19–24, 2010.

[538] Yamaoka, Y., “Pathogenesis of helicobacter pylori-related gastroduodenal diseases from molecular epidemiological studies,” *Gastroenterology research and practice*, vol. 2012, 2012.

[539] Yandell, M. D. and Majoros, W. H., “Genomics and natural language processing,” *Nature Reviews Genetics*, vol. 3, no. 8, pp. 601–610, 2002.

[540] Yang, H., Nepusz, T., and Paccanaro, A., “Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty,” *Bioinformatics*, vol. 28, no. 10, pp. 1383–1389, 2012.

[541] Yu, S., Van Vooren, S., Tranchevent, L.-C., De Moor, B., and Moreau, Y., “Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining,” *Bioinformatics*, vol. 24, no. 16, pp. i119–i125, 2008.

[542] Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., and Lazarus, R., “Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system,” *BMC medical informatics and decision making*, vol. 6, no. 1, p. 30, 2006.

[543] Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., and others, “Structure-based prediction of protein-protein interactions on a genome-wide scale,” *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[544] Zhang, R., Wang, Y., and Melton, G. B., *Medical Applications of Artificial Intelligence*, ch. 23. Natural Language Processing in in Medicine. CRC Press, 2013.

[545] Zhang, W., Sun, F., and Jiang, R., “Integrating multiple protein-protein interaction networks to prioritize disease genes: a bayesian regression approach,” *BMC bioinformatics*, vol. 12, no. Suppl 1, p. S11, 2011.

[546] Zhang, Z., Bajic, V. B., Yu, J., Cheung, K.-H., and Townsend, J. P., “Data integration in bioinformatics: current efforts and challenges,” *Bioinformatics-Trends and Methodologies*, pp. 41–56, 2011.

[547] Zhou, Z., Wang, Y., and Gu, J., "A new model of information content for semantic similarity in wordnet," in *Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on*, vol. 3, pp. 85–89, IEEE, 2008.

[548] Zikos, D., "Data issues for clinical-administrative decision making in healthcare,"

[549] Zimmerman, J. E., Kramer, A. A., McNair, D. S., and Malila, F. M., "Acute physiology and chronic health evaluation (apache) iv: Hospital mortality assessment for today's critically ill patients*," *Critical care medicine*, vol. 34, no. 5, pp. 1297–1310, 2006.