UNIVERSITY OF CINCINNATI

Date: 10/29/2007

I, Gudivada Ranga Chandra, hereby submit this work as part of the requirements for the degree of:

Doctorate of Philosophy

in:
Biomedical Engineering

It is entitled:

Discovery and Prioritization of Biological Entities
Underlying Complex Disorders by Phenome - Genome
Network Integration

This work and its defense approved by:

Chair: Bruce J Aronow PhD

Eric K. Neumann PhD

Marepalli Rao PhD
Discovery and Prioritization of Biological Entities Underlying Complex Disorders by Phenome - Genome Network Integration

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

In the Department of Biomedical Engineering

of the College of Engineering

By

Ranga Chandra Gudivada

B.Tech, Electronics & Instrumentation

Nagarjuna University, India, August 2001

Committee Chair: Bruce J Aronow PhD
Abstract

An important goal for biomedical research is to elucidate causal and modifier networks of human disease. While integrative functional genomics approaches have shown success in the identification of biological modules associated with normal and disease states, a critical bottleneck is representing knowledge capable of encompassing asserted or derivable causality mechanisms. Both single gene and more complex multifactorial diseases often exhibit several phenotypes and a variety of approaches suggest that phenotypic similarity between diseases can be a reflection of shared activities of common biological modules composed of interacting or functionally related genes. Thus, analyzing the overlaps and interrelationships of clinical manifestations of a series of related diseases may provide a window into the complex biological modules that lead to a disease phenotype. In order to evaluate our hypothesis, we are developing a systematic and formal approach to extract phenotypic information present in textual form within Online Mendelian Inheritance in Man (OMIM) and Syndrome DB databases to construct a disease–clinical phenotypic feature matrix to be used by various clustering procedures to find similarity between diseases. Our objective is to demonstrate relationships detectable across a range of disease concept types modeled in UMLS to analyze the detectable clinical overlaps of several Cardiovascular Syndromes (CVS) in OMIM in order to find the associations between phenotypic clusters and the functions of underlying genes and pathways.

Most of the current biomedical knowledge is spread across different databases in different formats and mining these datasets leads to large and unmanageable results. Semantic Web principles and standards provide an ideal platform to integrate such heterogeneous information and could allow the detection of implicit relations and the formulation of interesting hypotheses. We implemented a page-ranking algorithm onto Semantic Web to prioritize biological entities for their relative contribution and relevance which can be combined with this clustering approach. In this way, disease–gene, disease–pathway or disease–process relationships could be prioritized by mining a phenome–genome framework that not
only discovers but also determines the importance of the resources by making queries of higher order relationships of multi-dimensional data that reflect the feature complexity of diseases.
Acknowledgments

First I am deeply indebted to my supervisor, Dr. Bruce Aronow, for providing me with supervision, motivation and encouragement throughout my graduate studies. His insight, breadth of knowledge and enthusiasm has been invaluable. He has had the most profound influence on the way I think (and write!) about science and my approach to research in general. Without his care, supervision and friendship, I would not be able to complete this work. I feel privileged to have studied under a scientist of his stature. I am also grateful to my other supervisor, Dr. Eric Neumann, for his valuable suggestions and introducing me to the wonderful world of Semantic Web. His help has been integral to the success of this work. My deepest gratitude goes to Dr. Marepalli Rao for his guidance and assistance with this dissertation especially in validating my final results. Many thanks also go to other faculty at Department of Biomedical Informatics especially, Dr. Anil Jegga and Dr. Jason Lu for their help.

It has been an honor to be part of the Biomedical Informatics Department at Cincinnati Children’s Research Foundation and above all, I am fortunate to have labmets that have provided so much support, encouragements and love during my Ph.D. years and otherwise. Finally, I am indebted to my parents and brother for their affection, love and understanding all through my life.
# Table of Contents

Abstract..................................................................................................................... i

Acknowledgments .................................................................................................. iv

List of Figures...........................................................................................................3

Chapter 1: Introduction ..........................................................................................7
  Objectives ............................................................................................................... 8

Chapter 2: Background and Significance............................................................10
  Mendelian Disorders.............................................................................................. 10
  Complex Disorders ............................................................................................... 11
  Biomedical Informatics Approach.......................................................................... 11
  Phenome Networks – Genome Networks .............................................................. 12

Chapter 3: Bottlenecks and Current approaches in Biomedical Informatics .16
  Data Integration and Integrative Genomics ......................................................... 16
  Knowledge Representation ...................................................................................... 17

Chapter 4: Semantic Web for Life Sciences.......................................................20
  Semantic Web standards ...................................................................................... 20
  Data Integration on Semantic Web ....................................................................... 21
  Knowledge Representation utilizing Semantic Web .............................................. 22

Chapter 5: Creating and Validating Human Phenome groups .........................24
  Introduction ......................................................................................................... 24
  Methods.................................................................................................................. 26
  Results and Validation ......................................................................................... 33
  Case Studies .......................................................................................................... 38
  Discussion & Conclusion ...................................................................................... 39

Chapter 6: Construction and Validation of Genome Bio-RDF Network .......41
List of Figures

Figure 1: Genome – phenome network integration [12]............................................................................. 14

Figure 2: Semantic Web stack [38]........................................................................................................... 22

Figure 3: Data integration in Semantic Web using triples.......................................................................... 23

Figure 4: Shows the associations of Disease1, 2, 3 to their clinical features. Here, columns are UMLS concept identifiers (CUI) associated with clinical features .................................................... 31

Figure 5: Hierarchical relations present among CUIs of clinical features from above Phenomatrix. Relations are computed from MRHIER table at UMLS. It can be observed that ‘Heart Failure’ is the top most hypernym compared to all the other features.................................................. 32

Figure 6: a) Mapping all the diseases to the top most hypernym instead to associated hyponyms. b) UMLS concept identifiers and associated descriptions for the example.................................................. 32

Figure 7 a) Plot of Eigen value versus dimension b) Plot of normalized cumulative sum versus dimension........................................................................................................................................ 34

Figure 8: Phenotype similarity versus gene annotation similarities (a) proteins associated with similar phenotypes and sharing atleast one CDD domain (b) genes associated with similar phenotypes and sharing three or more GOA at the sixth or more detailed level. The average signal of 10 randomized phenomaps is at the lowest level. Disallow same gene analysis skips the disease pairs having same implicated gene................................................................................... 36

Figure 9: Marfan and Obesity disease groups confirm from literature findings......................................... 38

Figure 10: Schema diagram. A) Test gene set is obtained from a locus identified by linkage analysis, or a differentially expressed gene set from a microarray experiment. B) Genome and Phenome databases considered to create Bio-RDF includes GO: Molecular Function, GO: Biological Process, pathway, Mammalian Phenotype, OMIM and Syndrome DB C) Each resource in the Bio-RDF graph is scored for its importance in the network D) By issuing a SPARQL query relevant to a disease gene set, prioritized genes are obtained after computing the score for each result........................................................................................................................................ 45

Figure 11: Text (TX) and Clinical Synopsis (CS) section from OMIM for Cardiomyopathy dilated 1D (OMIM # 601494). As shown, the TX section provides additional clinical features not mentioned in CS section for that particular disorder........................................................................................................ 47
Figure 12: Clinical Synopsis sections associated with disorders Arterial Tortuosity Syndrome (OMIM # 208050) and Mucolipidosis IIIA (OMIM # 252600). Illustrating how to overcome orthography by semantic normalization of clinical features to UMLS concepts.......................... 48

Figure 13: Portion of Bio-RDF for gene DMD based on the DCO ontology. The upper network is the ontology providing the required semantics for the lower RDF network consisting instance data................................................................................................................................................ 51

Figure 14: Kleinberg’s Authoritative and Hub Nodes................................................................. 54

Figure 15: a) Illustrating the significance of a gene associated with multiple pathways is considered more important compared to a pathway having multiple genes b) Assigning subjectivity and objectivity weights to the property ‘associatedPathway’ for the triple ‘gene – associatedPathway – Pathway’.................................................................................................... 55

Figure 16: Rank ROC curves for validation of integration - Each curve is generated based on rankings of the implicated gene out of 300 genes (on average) from the loci associated with 60 sample OMIM diseases. The 4 curves, as indicated by different colors are associated with sequential integration of different genome – phenome data sources. The data sources used to construct every ROC curve are indicated on the figure................................................................................................................. 55

Figure 17: Sequential addition of Genome – Phenome datasets improves SDBH gene ranking implicated in PARAGANGLIOMAS 4........................................................................................ 61

Figure 18: Rank ROC curves for validation a) each curve is generated based on rankings of the implicated gene out of 300 genes (on average) from the loci associated with 60 sample OMIM diseases including all data sources. The curves provide the comparison of overall ranking performance with and without semantic weights b) each curve is generated based on different combination of weights. The curve provides evidence of importance in giving more subjective weight for particular properties..................................................................................................... 62

Figure 19 Ranked pathways from various sources from the Bio-RDF associated with differentially expressed genes in human idiopathic dilated cardiomyopathy (DCM) [16] ............. 64

Figure 20: High-order View of Disease Card Ontology[124]. A simplified schematic representation of ontology classes used for integrating biological, pharmacological, and phenomical data .................................................................................................................................................................................................................................................. 68

Figure 21: RDF data models (a) Genome RDF model (b) Phenome RDF model.......................... 70

Figure 22: RDF resource representation differentiating multiple entities ................................... 71

Figure 23: N-ary relationship model to hold the phenomap of Type1 matrix ............................... 73
Figure 24: Anatomy of ORACLE – SPARQL query[131] .......................................................... 75

Figure 25: Real work flow model of creating, loading and forming Phenome – Genome network in the Oracle RDF store (Seamark Demonstration Image) ................................................................. 77

Figure 26: Computable sub-phenome network ........................................................................ 79

Figure 27: Computable sub-Genome network ........................................................................ 79

Figure 28: Computable sub-phenome – genome network ....................................................... 80

Figure 29: The total work flow from querying a phenotype group to ranking its associated biological entities ................................................................................................................................. 81

Figure 30: Two level genome network associated with each disease in the cluster ............... 82

Figure 31: Log – Log plot showing Scale- free nature of networks [137] ............................... 84

Figure 32: Log – Log plot depicting scale – free properties of Marfan syndrome group associated genome BioRDF network ........................................................................................................... 85

Figure 33: RDF network permutation procedure. Total degree for each ‘gene’ node is decomposed into its ‘typed degree’ belonging to type of edges. ‘Typed degrees’ are permuted in such a way to keep the total ‘type degree’ distribution constant. Objects nodes (pathways, GO terms) are randomized before adding to their specific property ................................................................. 86

Figure 34: Illustrating the scale – free nature of permuted networks (for Marfan Syndrome group) ............................................................................................................................................. 87

Figure 35: Information about Marfan phenome group curated from literature ................. 96

Figure 36: Predicted from clustering and ranking associated genome-BioRDF of the group .... 97

Figure 37: RDF graph associated with Marfan syndrome group with visual depiction of ranked entities ........................................................................................................................................ 98

Figure 38: Histogram of 1000 P-values resulting from correlation tests of real and permuted network ranks of genes relevant to Marfan syndrome group ............................................................. 99

Figure 39: Histogram of 1000 P-values resulting from correlation tests of real and permuted network ranks of pathways relevant to Marfan syndrome group ...................................................... 100
Figure 40: Proof of these particular entities are significant though they have less Chi-square value............................................................................................................................................ 104

Figure 41: Obesity disease group curated from literature [1, 5-9]....................................................... 106

Figure 42: RDF visualizer of ranked entities associated with Obesity group................................. 108

Figure 43: Histogram of 1000 P-values resulting from correlation tests of real and permuted network ranks of pathways relevant to Obesity phenotype group.............................................. 110

Figure 44: Proof of these particular GO-Biological processes associated with obesity group are significant though they have less Chi-square value......................................................... 116

Figure 45: Proof of these particular GO-Mol Functions associated with obesity group are significant though they have less Chi-square value................................................................. 116

Figure 46: Obesity associated graph depicting that ‘Angiotensin-converting enzyme 2 regulates heart function’ is ranked high due to its high scored two participating genes (AGT, AGTR1) from our data set. ........................................................................................................................ 118
Chapter 1: Introduction

The focus of disease-susceptibility gene discovery efforts is shifting from rare, monogenic conditions towards the common multi-factorial conditions that account for the majority of human illness and mortality. The susceptibility in these conditions is attributed to the effects of genetic variation at a number of different genes and their interaction with relevant environmental exposures. The expectation is that identification and characterization of the genes providing the inherited component of susceptibility will lead to substantial advances in our understanding of disease and, in turn, to improvements in diagnostic accuracy, prognostic precision, and the range and targeting of available therapeutic options (McCarthy et al., 2003)[10]. Though achieving this goal is relatively simple in single gene syndromes, more complex diseases often consist of multiple clinical phenotypes that might be the result of interactions among multiple genetic loci.

Although integrative genomics based approaches have been shown to be successful in understanding the underlying pathways and biological processes in normal and disease states, most of the current biomedical knowledge is spread across different databases in different formats. The very quantity, intricacy and diverse nature of data continue to pose a formidable challenge in dissecting etiology of complex disorders. Therefore it is crucial to represent knowledge and observations in a sufficiently standardized manner by linking multiple ontologies from various domains to permit inferences and assertions associating genotypes to phenotypes.

The objective of this study is to use a biomedical informatics approach, with genome to phenome scale semantic integrative analysis and knowledge representation as core aspects to accelerate the discovery of genes, pathways and other biological modules underlying complex
disorders. With the feasibility from the clinical point of view to dissect disease mechanisms[11-14], we hypothesize that “A framework that systematically links similarity-based network of syndrome phenotypes to functional-relatedness-based genome network will provide a knowledge model that facilitates discovery of novel genes, processes, or pathways underlying diseases with known or unknown etiology”.

Achieving this goal fundamentally requires integrating clinical and genomic databases and creating a “scalable semantic infrastructure”, that deals with its semantics and complexity, allowing integration and interoperability. Such an infrastructure requires semantic biological information, categorized in structured domains with biomedical-backed ontology framework as reference terminology.

Objectives

We plan to test our hypothesis and accomplish the goals of the proposed study by pursuing the following three objectives:

**Objective 1: To create and validate similarity-based syndrome phenotype networks by semantically integrating and clustering disease-centric data.**

a) Data extraction and aggregation relevant to diseases and clinical features

b) Integration and semantic normalization of disease centric clinical features from the previous step and prepare data sets for clustering

c) Survey of clustering algorithms to be applied on pheno-matrix (Matrix of syndromes as rows and clinical features as attributes)

d) Case studies and statistical validation of results
**Objective 2: To create and validate functional-relatedness-based network of genes using phenome-genome associations, interactome, pathways and gene ontologies (GO)**

a) Mining and integrating human phenome, interactome, GO, and pathway

b) Create a functional genome –BioRDF from the extracted data

c) Investigation and implementation of resource ranking algorithms on the genome - BioRDF

d) Case studies and statistical validation to test the efficacy of resource ranking algorithms

**Objective 3: To design, integrate, construct and validate phenome-genome networks developed in objectives 1, 2**

a) Design of RDF data models to convert and support integration of phenome and genome networks

b) Conversion, loading and construction of phenome – genome network

c) Implementation of ranking algorithm on to network

d) Design of work flow process model to initiate graph based network traversal methods, capable of posing richer and complex queries to effectively mine the biological entities underlying complex diseases and retrieve results across the cross –domain RDF network

e) Case studies and statistical validation of results
Chapter 2: Background and Significance

One of the principal goals of biomedical research is to elucidate the complex causal, modifier, and suppressor network of interactions from specific molecular pathways to whole organ systems underlying common human diseases. Though achieving this goal is relatively simple in single gene syndromes, more complex and rare diseases often consist of multiple clinical phenotypes that might be the result of interactions among multiple genetic loci[12]. Botsein & Risch [15] suggested that the disease genes discovered to date represent the easy ones, and the discovery of the remaining Mendelian and complex disease phenotypes will be difficult due to the rarity of the phenotypes, due to the heterogeneity or because of complex genetics where multiple genes and modifiers are contributing to a phenotype. The next two sections compare and contrast Mendelian with complex disorders.

**Mendelian Disorders**

The past few decades witnessed an explosion in the identification of genes associated with Mendelian diseases. Traditional approaches such as candidate gene approach and positional cloning via linkage analysis were the two prominent approaches utilized in discovering the disease genes underlying human diseases [12]. The candidate gene approach depends on prior knowledge about disease genes, such as tissues in which they are expressed or putative functional domains. Genes are prioritized using these clues and sequentially tested in association studies for segregating mutations and polymorphisms. Positional cloning doesn’t consider the prior knowledge regarding gene function. The studies are carried out in families with multiple affected members using microsatellite markers and other DNA polymorphisms. Alleles of the
markers that segregate with the disease help delineate a critical region where the probability of finding a disease gene is more [15].

**Complex Disorders**

Methods for studying complex phenotypes can be divided into two basic approaches, gene driven, which focus on certain genes in order to discover the phenotypes they influence and trait driven, which focuses on phenotypes and look to find the causative genes. Knockout models are the traditional method for proving and analyzing traits influenced by single genes. But complex phenotypes are affected by multiple, potentially unknown loci as well as epistatic relations among them. This requires more complicated, multivariate methods of analysis [11, 16].

**Biomedical Informatics Approach**

The post-genomic era gave an unprecedented opportunity for medicine and biology to come ever close. Traditionally, Medical informatics (MI) deals with computational methods applying to clinical medicine and Bioinformatics (BI) deals with the intersection of computer science and biology [17]. Advancements in both the fields are providing new opportunities to correlate genotype information with the expressed phenotype and propose novel avenues for new testable hypothesis. Abundance of genomic knowledge supports the theory that diseases should be understood by considering the complex interactions between genes and environmental factors that initiate pathological processes and define phenotype information [17, 18]. In the next few sections, a knowledge mining approach not only to discover underlying genes, disease pathways and interactions but also to rank and prioritize the significance of these entities underlying complex disorders is presented. This approach consists of constructing a framework to capture a biological knowledge model of two diverse domains functional networks in which one is a
disease network and the other is a gene network. The central idea behind this network approach has been adopted from Giallourakis C et al., 2004 [12].

**Phenome Networks – Genome Networks**

Understanding interactions among genes, from specific molecular pathways to the whole organ system should ultimately provide clinically useful insights into disease processes, including diseases that are influenced by multiple genetic loci. The hypothesis is based on the evidence provided from earlier research [19] that gene attributes correlate with phenotype sharing and explained by assuming that human phenotypes reflect disturbance of functional modules, more than an individual genes. Analyzing the molecular processes underlying phenotypically similar diseases may provide a window into these complex interactions[11, 13, 14, 16, 20, 21]. Therefore, a framework that systematically link clinical features to genomic elements will provide an excellent knowledge model to mine the entities underlying complex disorders (Figure: 1).

We consider two independent networks,

A) Network of functional relatedness of genes

B) Network of similarity-based syndrome phenotypes

In disease network, diseases can be related to each other on the basis of shared clinical symptoms, pathophysiology, etiology, anatomy and cellular endophenotypes. In the gene networks domain, genes can be related to each other in multiple ways as will be explained in the following section. The 1962 diseases in OMIM (statistics on August 29th from OMIM website) [22] which already have known gene associations will form a bridge between these two networks. These primary genes form the seeds to form gene networks, which can be associated
to the group of diseases sharing similar clinical features, in which the diseases of unknown etiology are part of. The gene networks associated with these disease genes provide an excellent candidates or modifiers with the query disease. The mapping between clinical features (disease networks or phenotype networks) and genes might become so robust that underlying genes, pathways or biological functions of a sporadic disease with unknown etiology might be identified by merely presenting the symptoms associated with the disease [12]. Earlier research showed that for similarity in human disease phenotypes showed a consistent association at multiple levels of gene annotation and their functions [11, 13, 19, 23, 24]. These networks can be mined to systematically prioritize genes, pathways, biological processes and molecular functions that can be tested individually or collectively for variation of human diseases. In the next two sections, Gene networks and Disease networks are considered in detail.

**Network of functional relatedness of genes**

Group of genes that cause a given phenotype tend to be linked at the biological levels as interacting proteins, as components of a multi protein complex or as steps in biochemical pathway. Therefore, data sets spanning molecular genomics data such as protein interactions and pathways to gene ontology annotations are integrated. Embedded within these vast datasets of information are correlations that weave together genes, genomic elements and clinical entities into functional networks [12]. By extrapolating the relations between the primary disease genes and other genes in the genomic databases we form the gene networks. Two genes might be related in this functional knowledge networks which include interactions and occurring in the same pathway, participate in same biological process and molecular function. Apart from these
Figure 1: Genome – phenome network integration [12]

relational types they also include co-occurrence relation from natural text such as Medline abstracts, gene – disease, anatomy, symptoms and other clinical features mined from UMLS (Unified Medical Language System) [25] including GeneRif records (Gene Reference into Function from NCBI). These networks will include well annotated OMIM genes (including 1962 genes linked to human diseases) as well as vast majority of genes and genomic elements where there is very little information. These large scale functional genomic experimental data sets provide complimentary views of gene function and facilitate more reliable grouping of genes based on shared roles in the cell that may not be apparent at the level of sequencing [12].
Network of similarity-based syndrome phenotypes

In order to construct the disease networks, diseases must be classified based on symptoms, anatomy and other clinical features. This can help in finding the hidden connections between complex diseases. Such connections between complex diseases reflect common biological pathways and biological functions that may become manifest in the form of co-morbidity. Phenotype grouping (disease clusters) reflects the modular nature of human disease genetics and general indicators of biological and functional interactions at the gene and protein levels. Traditionally diseases have been classified on the basis of pathophysiology or etiology. One of the approaches to classify diseases is by using cluster analysis and mapping the disease features to certain fixed concepts based on certain ontology. Unified Medical Language System (UMLS) provides an excellent standard framework to which the disease features can be mapped [25].

Achieving this goal requires integrating clinical databases with the genetic databases and also with other key challenges such as data aggregation, formal data representation protocols, sophisticated data mining and applying graph theoretical algorithms for ranking need to be addressed. The next few sections will discuss some of the techniques and practical approaches needed to boost sensitivity and specificity in constructing more faithful and functional knowledge networks spanning from genomics to clinical data in order to mine faithful relationships.
Chapter 3: Bottlenecks and Current approaches in Biomedical Informatics

Although Biomedical Informatics provides a probable solution to dissect complex phenomena underlying complex diseases, a number of constraints hinder the process of achieving the goal. They are general problems such as data integration, semantic mapping and knowledge representation to specific problems like clustering and ranking in the context of genome-phenome networks. In the next few sections, each problem and probable methodologies are described.

Data Integration and Integrative Genomics

Each technology interrogates different aspects of gene function. Integrating genome scale data sets provides complementary view points of a gene that could be useful, providing a more comprehensive description of functional gene networks. Genes and gene products do not function independently. They contribute to complex and interconnected pathways, networks and molecular systems. The understanding of these systems, their interactions and their properties will require information from several fields like genomics, proteomics, metabolomics or systematic phenotype profiles at the cell and organism level [26]. Each technology, especially the large-scale experimental data sets are extremely noisy and using information from single large-scale data set can lead to significant number of false positives. By integrating different types of functional genomics data can produce more reliable predictions with increased sensitivity and specificity for detecting true functional relationships unraveling of polygenetic disease causality and complex gene –environment interactions existent in disease pathogenesis
and causation. The benefits of integration are particularly valuable in prioritizing candidate human disease genes [27-33].

*Bio – Ontologies in Data Integration*

The very quantity, complexity and diverse nature of data continue to pose a formidable challenge to efforts to integrate and develop insights that can lead to new treatments by elucidating normal and disease physiology. Robust information management tools are therefore urgently needed to assist researchers in handling, navigating and integrating research findings. An ontological framework is required to perform information integration across heterogeneous data spanning from medicine to molecular biology which can allow multiple perspectives upon complex phenomena(such as diseases, associated risk factors, symptoms, pathological hallmarks, and so forth)[34]. This ontological framework will provide a way to query a disease and pool all the relevant and inferred information pertaining to that disease.

*Knowledge Representation*

Integrative genomics-based approaches have shown considerable success in the identification of linked pathways and biological processes in normal and disease states. However, a huge bottleneck for the accomplishment of deductive and predictive medicine is the failure to represent knowledge and observations in a sufficiently standardized manner to permit assertions of causality between disparate disease studies and biomedical knowledge. The issue of standardization becomes more important, as discoveries from biology and clinical medicine move from parallel to intersecting paths. Much of the biology works by applying prior
knowledge to an unknown entity, rather than the application of a set of axioms that will elicit knowledge [35]. Biological knowledge is inherently very complex capturing many to many relationships and often requires the addition of knowledge to specify and constrain the values held in the database. There is also a great deal of hierarchical knowledge in biology which needs to be integrated and represented with the existing molecular data and clinical data. Examples include anatomies, signal – transduction pathways and particularly phenotypic data such as clinical symptoms.

There are wide varied public and private biological databases and tools have been developed or been developed for genes, proteins, pathways, gene expression data, physiological processes, diseases, drugs, biomedical literature, etc. However, for biomedical researchers working on specific diseases, these resources end up being too generic. Data and information therefore needs to be embedded and represented in a disease context knowledge framework through data structures that reflect the organization of biological systems relevant to the disease.

**Bio – Ontologies in Knowledge Representation**

Ontology specifies at higher levels the classes of concepts that are relevant to the domain and the relations that exist between these classes. The concepts are defined by a set of assertions and by meaning that connect them to other concepts. Ontology captures the intrinsic conceptual structure of a domain and is a powerful medium to represent different knowledge dimensions and heart of knowledge representation. Bio – ontologies is an organizational framework of the concepts involved in biological entities and processes as well as medical knowledge in a system of hierarchical and associative relations that allows reasoning about biomedical knowledge. They are particularly helpful in particular at representing biological knowledge in a computer –
comprehensible way to improve knowledge discovery in life sciences databases and to use data mining techniques by providing a new semantic layer to the process and moving it from a data driven approach to a knowledge driven approach [36]. They describe the information in a way that is precise and formal enough to be manipulated by reasoning software and query tools. Inference across data sources is accomplished through computations that traverse a network of entities and relationships within ontology [37].
Chapter 4: Semantic Web for Life Sciences

The Semantic Web (SW) is a vision for the next generation web in which data from multiple sources described with rich semantics are integrated to enable human processing by humans as well as software agents [38]. One of the goals of the Semantic Web research is to incorporate a domain knowledge described with rich semantics into ontology that can be shared by many applications. Novel techniques can be utilized for effectively retrieving information and discovering hidden and implicit knowledge from the Semantic Web [39-47]. Bioinformatics is widely accepted as an important research area for the Semantic Web. Life Science resources are rich in data and knowledge and Semantic Web principles, standards and technologies provide an ideal platform to integrate such heterogeneous information and allow the detection of implicit relations embedded in clinical and genomic datasets. Semantic Web query languages such as SPARQL can be effectively used to mine the biological entities underlying complex diseases through richer and complex queries on this integrated data. In the next few sections, different Semantic Web languages and the advantages it provides in data integration, knowledge representation and reasoning with inference are discussed.

Semantic Web standards

Like every other data model, a Semantic Web data model needs a language to streamline the model into a format that is machine readable. The most well known languages are RDF (Resource Description Framework) [48], RDFS (RDF schema)[49] and OWL (Ontology Web Language) [50] (Figure:2). These three languages provide a unique format for the description and exchange of the semantics of web content. RDF provides a data model to represent resources and relations between resources and their values. A resource can be anything from a concrete to
abstract concept such as gene, disease, protein. Every resource in RDF is identified by a Uniform Resource Identifier (URI). The atomic element in a RDF is a triple of the form <Subject, Property, Object>. Here subjects are resources and objects can be resources or literals. Properties are objects that define binary relations in RDF model. RDF schema (RDFS) makes a RDF model more powerful by providing ontological views over RDF statements. For the purpose of efficient representation of ontologies on the Semantic Web OWL is the standard language recommended by W3C (World Wide Web Consortium). OWL is built on RDF schema and allows more details to be added to the resources. RDF schema imposes fairly loose constraints on the data model; OWL adds additional constraints that increase the accuracy of implementations of a given model.

**Data Integration on Semantic Web**

Most data integration projects fail. The common reason for the failure is the inability to extend the data model to incorporate new data, or inability to re-use data in ways that it was not originally intended. The main problem for integration efforts are lack of widely accepted standards for expressing the syntax and semantics of the data. RDF provides a very flexible model for adding new data to a data model and re-using data in ways that it was not originally intended [38]. The combined effect of global naming, universal data
structure and open world assumption is that resources in Semantic Web exist independently but can be readily linked with little pre-coordination. In an open world, only that which is explicitly stated is known—we cannot assume that something does not exist simply because it has not been stated. This means that new statements can be added without fear of breaking the data model, which happens too easily with existing schema mechanisms such as XML schema [47, 51, 52]. Figure: 3 presents an intuitive example to integrate data on Semantic Web using triples.

**Knowledge Representation utilizing Semantic Web**

The Semantic Web framework is a powerful approach to organize and represent different types of relationships between biological entities. The relationships range from physical interactions at molecular scale to complex observations such as effect of a drug on specific phenotype [53, 54]. Information stored as facts and relationships could support simulation of larger systems and
complex interactions which requires ontologies and logic which are supported by Semantic Web. OWL enhances the knowledge representation and inference capabilities of Semantic Web. Some of the RDF stores (e.g. Sesame[55] and Tucana[56]) do support OWL and provide reasoning facilities thus transforming them from a data store to a knowledge store.

Figure 3: Data integration in Semantic Web using triples
Introduction

One of the principal goals of biomedical research is to improve our understanding of how clinical phenotypes are enhanced or mitigated by underlying genotypic variations and mutations. Both single gene syndromes and more complex multifactorial diseases often exhibit several phenotypes and evidence from strain-based analyses strongly suggest that this can be driven by the interactions of multiple polymorphic genetic loci within and between distinguishable pathways and processes. A variety of approaches suggest that phenotypic similarity between diseases can be a reflection of shared activities of common biological modules composed of interacting or functionally related genes. Thus, analyzing the overlaps and interrelationships of clinical manifestations of a series of related diseases may provide a window into the complex biological modules that lead to pathophysiological processes to produce disease phenotype. However, few computational methods have been investigated for this purpose, in part due to intrinsic difficulties in accessing controlled clinical descriptions and its availability in structured form suitable for computational genome – wise analysis. In order to evaluate our hypothesis, we are developing a systematic and formal approach to extract phenotypic information present in textual form within Online Mendelian Inheritance in Man (OMIM) and Syndrome DB databases to construct a disease – clinical phenotypic feature matrix to be used by various clustering procedures to find similarity between diseases. Our objective is to demonstrate the roles of both generalization and specificity relationships detectable across a range of disease concept types modeled in UMLS to analyze the detectable clinical overlaps of several Cardiovascular
Syndromes (CVS) in OMIM in order to find the non-trivial associations between phenotypic clusters and the functions of underlying genes and pathways.

The modular nature of complex disorders can be attributed to their clinical feature overlap associated with mutations in different genes that are part of same biological module. Genes that are part of a functional module are connected at various biological levels as interacting partners, steps in a biochemical pathway, network of biological process or components of a multi-protein complex [13, 14, 19, 20]. Clustering diseases based on their shared clinical features provides an informational framework to analyze disease modularity and to explore underlying biological associations between various genotypic entities [11]. Indeed, this approach of logical grouping of genes by their associated phenotype clusters is referred as phenomics [19]. Phenomics approaches have potential applications in finding new disease genes that share functional annotations with known disease genes that are part of the same phenotype cluster and also to predict connectivity between apparently unrelated genes to the same functional module [14]. Nevertheless, there were few earlier attempts to systematically cluster diseases to gain insights into the molecular processes underlying them.

We used Online Mendelian Inheritance in Man (OMIM) (www.ncbi.nlm.nih.gov/omim/) [22] and National Library of Medicine – syndrome database (Syndrome DB) (http://www.nlm.nih.gov/archive/20061212/mesh/jablonski/syndrome_toc/toc_c.html) [57] as principal data sources for diseases and their corresponding clinical features. The phenotypic data presented in these data sources is not complete, unavailable in a well-ordered computable form and is not optimal to perform computations in forming accurate disease clusters. Therefore, this study only provides a proof of concept but certainly not a finished product. Our work has examined the possibility of using existing standard terminologies, metathesaurus[25] and text
mining tools[58, 59] to semantically normalize variations and also dimensionality reduction methods to overcome the complexity in dealing with large number of clinical features. We used similar validation methods presented in [19] to investigate correlations between disease similarity to their underlying gene sequence similarity and shared annotations at multiple levels including interactions, protein domains, gene ontology terms and pathways.

Although there are limitations with this approach, our analysis revealed there is a detectable correlation between phenotype similarities to multiple levels of gene annotations. In addition, we obtained some cogent results from case studies on grouping other diseases sharing clinical features with Marfan syndrome and Obesity. There are about 455 out of 977 CVS in OMIM having atleast one implicated gene and clinical feature. As, a pilot project we classified these 455 Cardiovascular Syndromes in OMIM forming a cardio-phenome system. We considered an OMIM disease as a cardiovascular syndrome (CVS) if it has at least one cardiovascular symptom mentioned in the clinical synopsis (CS) section or occurrence of terms such as “heart” or “cardiovascular” or “cardiac” in the free text section of OMIM.

**Methods**

**Data Sources**

Our data include both phenomic sources for forming disease clusters and genomic sources to validate the clusters by investigating phenotype – genotype correlations.

**Genomic Data Sources**

Human Gene Ontology- gene (GOA) annotations [60, 61], protein domains [62], protein sequences [63] and interactions [64] were downloaded from NCBI ftp site. Protein domains at
NCBI's Conserved Domain Database (CDD)[65] were not available for download and were parsed from the file ‘human.protein.gpff.gz’ [62] using Biojava software package[66]. Protein interactions at NCBI are imported from outside sources, such as BIND[67], BioGRID[68, 69] and HPRD[70, 71]. Gene-pathway annotations were compiled from KEGG[72, 73], BioCarta [74], BioCyc[75] and Reactome[76] databases.

Phenomic Data Sources

A total of 977 records were downloaded in XML format from OMIM by searching for terms ‘cardiovascular’ or “heart” or “cardiac” occurring in Clinical Synopsis (CS) section or Text Section (TX). Java XML parsers [77] were used to extract OMIM ID, disease name and the associated CS and TX sections from each OMIM record. OMIM ID and the corresponding gene associations were downloaded from NCBI Entrez Gene ftp site [78]. Syndrome DB is not available for download and java HTML parsers[79] were used to extract the data directly from their website. This database was developed by Stanley Jablonski and consists of 1600-2000 syndromes of congenital abnormalities known to be associated with mental retardation. Each entry has a ‘major features (MF) section’ (e.g. mouth and oral structures, abdomen and skin) similar to the CS section of OMIM. This database is web accessible at National Library of Medicine (NLM). A subset of 152 records having corresponding OMIM identifier and ‘cardiovascular system’ as one of the major clinical features were extracted. The entire clinical feature space encapsulates both clinical symptoms and affected anatomy. Clinical features under the categories such as “Inheritance” and “Molecular Basis” were eliminated. Nonspecific terms such as “syndrome” or “disease” or “disorder” were ignored.
**Matrix construction**

The pheno-matrix is a binary matrix in which all the rows are OMIM cardiovascular Syndromes (CVS) and columns are clinical features which comprise both clinical symptoms and affected anatomy. We assigned a value ‘1’ for the presence of clinical feature associated with an OMIM CVS and ‘0’ for no features. From the total of 977 CVS OMIM records, we took a subset, of 455 (46%) having at least one causative gene and a clinical feature. These 455 CVS are associated with 585 genes.

**Refining Clinical Features**

We performed a three step process to refine and reduce the clinical feature dimensional space.

- **a) Semantic Normalization**
- **b) Utilizing subsumption relations**
- **c) Principal Component Analysis (PCA)**

**a) Semantic Normalization**

The TX & CS sections of OMIM and MF section of Syndrome db are presented as loosely defined free textual descriptions. There is inconsistency in the use of clinical feature terms both semantically (e.g. increased sweating and diaphoresis) and syntactically (e.g. neonatal hypotonia and hypotonia, neonatal). In order to overcome these limitations, we have chosen to directly map these terms to Unified Medical Language System (UMLS) [25] concepts (CUI) thus performing Semantic Normalization. The corresponding UMLS concepts for the clinical features present in CS section of OMIM and MF section of Syndrome DB were mapped using MetaMap[58], a NLP (Natural Language Processing) tool which takes free text from biomedical domain and
maps noun phrases to a potential list of matching concepts from UMLS metathesaurus. We used an online version of metamap programme, available as part of Semantic Knowledge Representation project (SKR) [80], which aims to provide a framework for exploiting UMLS knowledge resources for NLP.

The extracted clinical features from CS & MF sections were uploaded into the metamap batch mode module and a java script was written to parse the results. The parser extracts score for each match, original textual phrase, mapped Concept Unique Identifiers (CUI’s) and the semantic type it belongs to from the list of final candidate mappings. To avoid erroneous mappings, UMLS Semantic Network is used to restrict the mappings belonging only to semantic types under ‘disorders’ and “anatomy” semantic group. These sets are further refined between scores ranging from 570 to 1000 and after careful manual curation, incorrectly assigned concepts were eliminated. SKR-Metamap works well for short phrases and gives exceptions while handling the TX section of OMIM as it contains large sections of free text as opposed to small phrases in CS. We used GATE toolbox (General Architecture for Text Engineering [59], produced at Sheffield University. GATE is a general purpose text engineering system, whose modular and flexible design allows us to use it to create a more specialized biological IE system. In our case, we used GATE for clinical feature entity recognition in the TX section of OMIM using gazetteers, an important component of GATE holding a list of members of a particular category. Here, the input to gazetteers is a list of clinical feature keywords supplied from UMLS concepts belonging to ‘disorders’ and ‘anatomy’ semantic groups. For each concept belonging to the above category, preferred names and synonyms were extracted and supplied to the gazetteers. Once, GATE scans through each OMIM TX section and identifies the clinical features matching
to the keywords present in the gazetteers, a post-processing step is performed to find the appropriate UMLS concepts for the extracted clinical features. Table 5-1 provides the statistics before and after performing semantic normalization of the clinical features. The advantage of using UMLS concepts instead of raw clinical features from unstructured text extremely reduced the total clinical feature space by around 50%.

<table>
<thead>
<tr>
<th>Clinical Features</th>
<th>Total Extracted Features</th>
<th>Total Features After Semantic Normalization</th>
<th>% Of Clinical Feature Reduction After Semantic Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Symptoms</td>
<td>16979</td>
<td>8504</td>
<td>50.08%</td>
</tr>
<tr>
<td>Affected Anatomy</td>
<td>8062</td>
<td>3364</td>
<td>41.7%</td>
</tr>
</tbody>
</table>

Table 5-1: Total clinical features extracted from OMIM and Syndrome DB before and after performing semantic normalization

b) Utilizing subsumption relations

UMLS Metathesaurus is a comprehensive database containing terms from more than 100 various source vocabularies, which are mostly biomedical terminologies having hierarchical relations (parent/child) providing surrogate subsumption relations (is a, subclass of). MRHIER table from UMLS provides the required hierarchical relations between UMLS concepts. As all the clinical features are mapped to UMLS concepts, we explored ways to further reduce clinical feature space utilizing the subsumption relations in MRHIER table. Figures 4, 5 and 6 provide a hypothetical example, where we scanned the clinical feature set to find the top most hypernym and link all the diseases to that specific hypernym instead of the hyponyms associated with the
disease. By reducing the clinical features using subsumption relations and also ignoring the clinical features associated with only one particular disorder, the final clinical feature set contained 1916 features.

c) **Principal Component Analysis**

In order to find the best disease clusters, we considered each OMIM CVS as a datum. Since the number of diseases (455 CVS) is much smaller than the number of the feature dimensions (1916 clinical features) and the features are all binary, the data contains a great deal of redundancy. In order to detect meaningful underlying dimensions from our high dimensional dataset, we had chosen Principal Component Analysis as it efficiently reduces high dimensional data set into low dimensional map. PCA is a linear subspace identification method, which finds the correlative embedding that maximizes the projected variance of the original data. Most of the similar disease features are captured

<table>
<thead>
<tr>
<th></th>
<th>C0034072</th>
<th>C0235527</th>
<th>C0221045</th>
<th>C0264722</th>
<th>C1504382</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease 1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disease 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disease 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 4:** Shows the associations of Disease 1, 2, 3 to their clinical features. Here, columns are UMLS concept identifiers (CUI) associated with clinical features
Figure 5: Hierarchical relations present among CUIs of clinical features from above Phenomatrix. Relations are computed from MRHIER table at UMLS. It can be observed that ‘Heart Failure’ is the top most hypernym compared to all the other features.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Heart Failure [C0018801]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease 1</td>
<td>1</td>
</tr>
<tr>
<td>Disease 2</td>
<td>1</td>
</tr>
<tr>
<td>Disease 3</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UMLS Concept Identiﬁers (CUI)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0018801</td>
<td>Heart failure</td>
</tr>
<tr>
<td>C0264722</td>
<td>Chronic congestive heart failure</td>
</tr>
<tr>
<td>C0235527</td>
<td>Right ventricular failure</td>
</tr>
<tr>
<td>C0221045</td>
<td>High output heart failure</td>
</tr>
<tr>
<td>C0264722</td>
<td>Chronic congestive heart failure</td>
</tr>
<tr>
<td>C1504382</td>
<td>Pulmonary artery wall hypertrophy</td>
</tr>
</tbody>
</table>

Figure 6: a) Mapping all the diseases to the top most hypernym instead to associated hyponyms. b) UMLS concept identifiers and associated descriptions for the example.

by the first set of components (reduced-dimension), while including all components raises the captured variance fully to 100% (spectral energy). Note that the PCA feature is not necessarily binary though the original data is. To analyze the energy distribution of the PCA results, we calculated and arranged the eigen values of the data space from largest to smallest. Figure 7-a gives the plot of eigen-value versus dimension. From the normalized cumulative sum (Figure 7-b), we see that only 68% spectrum energy for the top 82 principal components are good enough
to reproduce the classifications corroborating the evidences extracted from the literature. Selecting the top 82 principal components (from earlier 1916 clinical features obtained after using subsumption relations) clearly explicates an effective dimensionality reduction. Matlab 7.0 (http://www.mathworks.com/), a popular Mathematical toolset was used to perform PCA and to generate the plot.

Similarity measure and Clustering

The similarity between two CVS is calculated by measuring the cosine of angle between the associated clinical features vectors obtained after PCA. Hierarchical Clustering was performed on the resulting 455 x 455 phenomap (distance matrix obtained after applying cosine distance on the 455 x 82 pheno-matrix, where 82 are the top principal components). ‘R’ statistical software package [81] was used for this analysis.

Results and Validation

The average of 10 randomized phenomaps was used as a control for background signal. Clinical feature vectors were randomly permuted using Fisher – Yates shuffling [82] before applying PCA and we considered the top principal components associated with 68% of spectral energy as like original matrix. We used similar validation approaches as[19] , to test the initial hypothesis that similarities at the phenotypic level correlate to similarities in gene/protein function. Though we performed a thorough analysis correlating phenotype overlap to multiple levels of gene annotation similarities, due to space limitations, here we present the results only for co-occurring domains and GO annotations.
**Phenotype similarity – protein sequence similarity correlations**

The 455 CVS are associated with 869 protein coding sequences (from 585 genes) extracted from NCBI database. A pair wise 869 X 869 protein sequence alignment is performed using Smith-Waterman Analysis algorithm (Blosum-62). Sequence pairs with an alignment similarity percentage better than 75% were considered similar. JAligner [83], an open source Java implementation of the Smith-Waterman algorithm was used to perform this analysis. Figure 8-b shows the significant sequence alignment similarities as a function of the phenotype similarity scores. The percentage of CVS pairs for which the causative proteins sequences are similar increases with increasing phenotype similarity score from 0.45% to 19%. We further disallowed the same gene associated with two similar CVS and from the Figure 8-b approximately half of these are owing to different mutations in the same gene causing the similar phenotype.

**Figure 7**

(a) Plot of Eigen value versus dimension  
(b) Plot of normalized cumulative sum versus dimension
**Phenotype similarity – Domain co-occurrence**

Proteins share functional domains and a mutation occurring in a shared domain might disrupt a specific biological process or a pathway leading to similar phenotypes. Figure 8-c shows the percentage of protein pairs that share a CDD domain as a function of the phenotypic similarity scores. The percentage of shared domains increases with increasing phenotype similarity score from 0.3% to 15%. For instance, ‘Cardiomyopathy, Dilated, 1E’ [OMIM: 601154] is caused by a mutation in SCN5A [NCBI GENE ID: 6331] and shares phenotypic characteristics with ‘Jervell and Lange-Nielsen Syndrome’ [OMIM: 220400] that is caused by a mutation in KCNQ1 [NCBI GENE ID: 3784]. These two proteins have a common ‘sodium ion transporter’ domain [CDD: 70001].

**Phenotypic similarity – shared interacting partners**

We also examined whether the phenotypic similarity correlates with shared interacting partners. Figure 8-d illustrates that percentage of shared interacting partners associated with primary genes implicated with the CVS increase with phenotypic similarity. The majority of shared interacting partners (23.6%) can be seen at 0.8 – 0.9 phenotypic similarity pair bin.
Figure 8: Phenotype similarity versus gene annotation similarities (a) proteins associated with similar phenotypes and sharing at least one CDD domain (b) genes associated with similar phenotypes and sharing three or more GOA at the sixth or more detailed level. The average signal of 10 randomized phenomaps is at the lowest level. Disallow same gene analysis skips the disease pairs having same implicated gene.

**Phenotype similarity – Gene Ontology correlations**

To explore possible functional relations between genes associated with overlapping CVS, we compared GOA. Similar to the earlier work [19], we defined GO similarity by the sharing of at least three GOA at the sixth or more detailed GO level. From figure 8-e the percentage of CVS
pairs that share three or more GOA increased (from 1.15% to 33.33%) as a function of the phenotypic similarity. The signal we find is well above the average of 10 randomized matrices (~2%) over all bins.

*Phenotype similarity – shared pathway correlations*

Genes from a same biochemical pathway together perform as a single biological module. Mutations in these genes may thus lead to the same or similar phenotypes, potentially resulting in the phenomenon of syndrome families [19, 20]. As explained in the methods section, we pooled pathways from several pathway data sources. From figure 8-f it’s clearly visible that phenotypic pairs having high similarity scores (from 0.7 to 1.0) are correlated with sharing high percentage of pathways compared with the less similar phenotypic pairs.
Figure 9: Marfan and Obesity disease groups confirm from literature findings

Case Studies

To test the feasibility of our approach we compiled related disorders that share clinical features with Marfan syndrome and Obesity from the biomedical literature. Marfan syndrome is a disorder of connective tissue, the tissue that strengthens the body's structures[3, 84-86]. From CS section of OMIM, Marfan syndrome [OMIM: 154700] has several cardiovascular clinical features such as congestive heart failure, mitral regurgitation and aortic root dilatation. Obesity is becoming a global epidemic in both children and adults, and it is associated with numerous co-morbidities such as hypertension, congestive heart failure, coronary heart disease and Type2 diabetes[1, 87-90]. Obesity [OMIM: 601665] was considered as CVS as the TX section of OMIM includes terms such as ‘cardiovascular’ and ‘heart’. Figure 9 validates that most of the
diseases clustered with obesity and Marfan syndrome was actually found related in literature [30, 31].

**Discussion & Conclusion**

Though our work is closely related to [13, 19], we did deviate in several ways by using Syndrome DB in addition to OMIM and further, reducing clinical feature dimension space in utilizing subsumption relations from UMLS and implementing PCA were novel. We were not able to compare our results with the earlier work (analyzed 1653 OMIM phenotypes) as here we concentrated on subset of OMIM diseases (455 CVS). Genomics has given researchers a collective set of biomedical resources that serve as a molecular catalogue for associating symptomatic and mechanistic evidence of diseases and disorders. This is unprecedented in the history of medicine, and demarcates the beginning of increasing knowledge in therapeutics areas. However, at this juncture in time, we have many strands of information of varying forms and reliability that can increase our overall understanding if we know how best to utilize them.

Indeed, the recent published work by the The Wellcome Trust Case Control Consortium[91] illustrates that genome wide approaches require large sampling of molecular clinical data (polymorphisms, symptoms), and yield many more gene associations than previously obtained from earlier limited diseases studies. We have shown how semantically rich information of symptoms can be used to relate their causative diseases to underlying genetic components. In previous work[92], we have shown how Semantic Web standards can be used to aggregates broad sets of data, and prioritizing them for relative contribution and relevance using a page-ranking approach, and which we are currently combining with this clustering approach.
This differs substantially by work of others, such as Goh et al[93], who used OMIM information directly to link genes with diseases. Our approach has the advantage that we utilize the breadth of symptomatic evidence to establish possible associations between different diseases, thereby not relying solely on our current limited knowledge of the genetic basis. We see our approach being able to accommodate the inherent noise in the information sets, whether it comes from improper classifications, speculative hypotheses, or faulty reasoning. As more evidence and interpretations get compiled, we anticipate broad semantic analysis becoming more robust and yielding more insights. We believe the use of diverse sets of evidence and hypotheses will greatly advance the set of testable models for overlapping diseases mechanisms. This will consequently have a direct impact in the development of both new and re-directed therapeutic applications. The key requirement we propose is that the application of semantics in data sets and curated interpretations help to manage and analyze the vast biomedical resources being generated, and will become an indispensable tool for biomedical researchers.
Chapter 6: Construction and Validation of Genome Bio-RDF Network

Introduction

Most common chronic diseases are multifactorial and characteristically involve the responses and influences of susceptibility and modifier genes that are subject to environmental factors. These interactions, mechanisms and phenotypic consequences can be richly represented using scale-free networks with semantically definable nodes and edges. Genomic studies using linkage analyses detect quantitative trait loci that encompass a large number of disease candidate genes. Similarly, transcriptomic studies using differential gene expression profiling generate hundreds of potential disease candidate genes that themselves may not include genetically variant genes that are responsible for the expression pattern signature. Hypothesizing that the majority of disease causal genes are biochemically known to play functionally important roles and whose mutations produce clinical features similar to the disease under study, we reasoned that an integrative genomics-phenomics approach utilizing the available annotation and clinical phenotypes derived from human and mouse gene orthologs could expedite disease candidate gene identification and prioritization. To approach the problem of inferring likely causality roles, we generated Semantic Web methods-based network data structures, and performed centrality analyses to rank genes according to model-driven semantic relationships. Our results indicate that Semantic Web approaches enable systematic leveraging of implicit relations hitherto embedded among large datasets and can greatly facilitate identification of centrality elements that can lead to specific hypotheses and new insights.
The identification of genes responsible for causing or preventing human disease provides critical knowledge of underlying pathophysiological mechanisms and is essential for developing new diagnostics and therapeutics. Traditional approaches such as positional cloning and candidate gene analyses, as well as modern methodologies such as gene expression profiling tend to fail to discover genes underlying diseases. Quantitative trait loci intervals identified by positional cloning usually embed a few dozens. Prioritizing candidates within these lists tends to be difficult, thus techniques and tools to identify key candidates from gene lists generated by disease process-associated gene discovery methods would be very desirable. Moreover, the demonstration of successful methods for the identification of disease-critical genes would also serve to validate specific computational approaches useful for knowledge representation and inference for the improvement of human health.

Disease gene discovery has been shown to be accelerated by applying aggregative computational methodologies on integrated data sets generated from genome-scale experiments [32]. Integrating diverse functional genomic data has several advantages as described by Giallourakis et al [12]. First, a more comprehensive description of functional gene networks can be formed by combining complimentary view-points generated from interrogation of diverse aspects of gene function from different technologies. Second, data integration reduces noise associated with each experimental limitation, thus increases sensitivity and specificity to detect true functional relationships which results in less number of false positives. However, large scale data aggregation efforts tend to be manual and lack sufficient semantic abstraction to allow for mechanistic generalizations.

Several gene prioritization methods have been developed [23, 27, 31-33, 94-97]. Some of them [23, 27, 31, 32, 94]use training gene sets to prioritize candidate test genes based on their
similarity with the training properties obtained from the reference set. One significant drawback in these methods is dependence on training set genes, because in many practical situations relevant training sets are not available and results may also vary depending on differing approaches used to delineate the particular training set used. There are few other methods [33, 95, 96] which do not require any training set in prioritizing candidate test genes but their potential is limited by accessing only few data sources. Here, for the first time we utilized Semantic Web (SW) [38] standards and techniques for hunting human disease genes. Resource Description Framework (RDF) (www.w3.org/RDF/) and Ontology Web Language (OWL) (www.w3.org/2004/OWL/) are used to integrate genomic and phenomic annotations associated with candidate gene set. The resulting Bio-RDF is a conventional directed acyclic graph (DAG) and centrality analysis is applied to score the elements in the network based on their importance within network structure. Scoring of each gene depends on the functional importance obtained from the genome data combined with clinical features it’s sharing with related diseases obtained from phenomic data. Centrality measures are calculated from a modified version [98] of Kleinberg algorithm [99] extended for SW. Central elements of biological networks are found to be functionally essential for viability and can lead to new insights to generate new hypotheses [100]. Apart from Kleinberg Authoritative Scores, there are several other centrality measurements, such as Google PageRank [101], Centrality Indices [102] and C-F closeness[103]. While SW querying languages do not naturally rank the retrieved results from RDF graphs, we have adapted a technique described by M. Sougata et al [40, 98, 104] for domain specific ranking to rank the retrieved genes from Bio-RDF using SPARQL (http://www.w3.org/TR/rdf-sparql-query/).
Our approach has enabled for the first time to utilize the combination of mouse phenotypes and human disease clinical features apart from GO and pathways in their prioritization approach. Our method doesn’t use any training data set, but extends the earlier hypothesis that majority of disease causal genes are functionally important and share clinical features with related diseases[32, 33]. We reasoned that an integrative genomics-phenomics approach utilizing the available human gene annotations including human and mouse phenotype data will expedite disease candidate gene identification and prioritization. In the current study we focused on the cardiovascular system diseases (CVD). We tested this hypothesis by prioritizing (a) genes from the recently reported cardiomyopathy susceptibility loci (chromosomes 7p12.1-7q21 [105] and (b) genes differentially expressed in dilated cardiomyopathy [106].

**Methods**

**Data Sources**

We used both genomic and phenomic data sources to prioritize gene candidates (See Figure 10).

**Genomic Data Sources**


2) Gene-pathway annotations were compiled from KEGG [73], BioCarta (http://www.biocarta.com/), BioCyc [75] and Reactome [76]. 4772 human genes had at least one pathway association (a total of 672 pathways).
Figure 10: Schema diagram. A) Test gene set is obtained from a locus identified by linkage analysis, or a differentially expressed gene set from a microarray experiment. B) Genome and Phenome databases considered to create Bio-RDF includes GO: Molecular Function, GO: Biological Process, pathway, Mammalian Phenotype, OMIM and Syndrome DB. C) Each resource in the Bio-RDF graph is scored for its importance in the network. D) By issuing a SPARQL query relevant to a disease gene set, prioritized genes are obtained after computing the score for each result.

Phenomic Data Sources

- Mammalian Phenotype (MP) ontology [107] and mouse gene phenotype annotations and the corresponding orthologous human genes were downloaded from Mouse Genome Informatics (MGI) website (http://www.informatics.jax.org). This data set contained 4127 human genes annotated with 4066 mouse phenotypes.
- A total of 977 records (423 have at least one implicated gene) were downloaded in XML format from OMIM [108] by searching for terms ‘cardiovascular’ or “heart” or “cardiac”
occurring in Clinical Synopsis (CS) section or Text Section (TX). Java XML parsers (http://xerces.apache.org/xerces-j/) were used to extract OMIM ID, disease name and the associated CS and TX sections from each OMIM record. We considered parsing each TX section of OMIM record as it provides additional clinical features to the ones available from CS section, which is evident from Figure 11. The entire clinical feature space encapsulates both clinical symptoms and affected anatomy. Clinical features under the categories such as “Inheritance” and “Molecular Basis” were eliminated. Nonspecific terms such as “syndrome” or “disease” or “disorder” were ignored. OMIM ID and the corresponding gene associations were downloaded from NCBI Entrez Gene ftp site (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/ mim2gene).

- The Multiple Congenital Anomaly/Mental Retardation database (Syndrome db) was not available for download and java HTML scripts were used to extract the data directly from the website. This database was developed by Stanley Jablonski [109] and consists of structured descriptions of approximately 700 out of the 1600-2000 syndromes of congenital abnormalities known to be associated with mental retardation. Each entry has a ‘major features (MF) section’ (e.g. mouth and oral structures, abdomen and skin) similar to the CS section of OMIM. This database is web accessible at National Library of Medicine (NLM). A subset of 152 records having corresponding OMIM identifier and ‘cardiovascular system’ as one of the major clinical features were extracted.
**Text Section**

Figure 11: Text (TX) and Clinical Synopsis (CS) section from OMIM for Cardiomyopathy dilated 1D (OMIM # 601494). As shown, the TX section provides additional clinical features not mentioned in CS section for that particular disorder.
Figure 12: Clinical Synopsis sections associated with disorders Arterial Tortuosity Syndrome (OMIM # 208050) and Mucolipidosis IIIA (OMIM # 252600). Illustrating how to overcome orthography by semantic normalization of clinical features to UMLS concepts.

Mapping Clinical Features to Find UMLS Concepts

OMIM ID’s and the corresponding features from CS section are parsed using java XML scripts from the downloaded XML files. The CS section of OMIM and MF section of Syndrome db are presented as loosely defined free textual descriptions. There is inconsistency in the use of clinical feature terms both semantically (e.g. increased sweating and diaphoresis) and syntactically (e.g. neonatal hypotonia and hypotonia, neonatal). In order to overcome these limitations, we have chosen to directly map these terms to Unified Medical Language System (UMLS) (http://umlsks.nlm.nih.gov).
Table 6-1: Total clinical features extracted from OMIM and Syndrome db before and after performing semantic normalization (mapping to UMLS concepts).

Concepts using MetaMap [58]. It’s a NLP (Natural Language Processing) tool which takes free text from biomedical domain and maps noun phrases to a potential list of matching concepts from UMLS metathesaurus. Figure 12 provides an example to overcome orthography (spelling variants) problem inherent in clinical terms present in OMIM records by mapping to UMLS concepts. We used an online version of metamap programme, available as part of Semantic Knowledge Representation project (SKR) (http://skr.nlm.nih.gov/), which aims to provide a framework for exploiting UMLS knowledge resources for NLP.

The extracted clinical features were uploaded into the metamap batch mode module and a java script was written to parse the results. The parser extracts score for each match, original textual phrase, mapped Concept Unique Identifiers (CUI’s) and the semantic type it belongs to from the list of final candidate mappings. To avoid erroneous mappings, UMLS Semantic Network is used to restrict the mappings belonging only to semantic types under ‘Disorders’ semantic group. These sets are further refined between scores ranging from 570 to 1000 and after careful manual curation, incorrectly assigned concepts were eliminated. SKR-Metamap works well for short phrases and gives exceptions while handling the TX section of OMIM as it contains large sections of free text as opposed to small phrases in CS. We used GATE...
toolbox[59] (General Architecture for Text Engineering), produced at Sheffield University. GATE is a general purpose text engineering system, whose modular and flexible design allows us to use it to create a more specialized biological IE system. In our case, we used GATE for clinical feature entity recognition in the TX section of OMIM using gazetteers, an important component of GATE holding a list of members of a particular category. Here, the input to gazetteers is a list of clinical feature keywords supplied from UMLS concepts belonging to ‘disorders’ and ‘anatomy’ semantic groups. For each concept belonging to the above category, preferred names and synonyms were extracted and supplied to the gazetteers. Once, GATE scans through each OMIM TX section and identifies the clinical features matching to the keywords present in the gazetteers, a post-processing step is performed to find the appropriate UMLS concepts for the extracted clinical features. Table 6-1 provides the statistics before and after performing semantic normalization of the clinical features. The advantage of using UMLS concepts instead of raw clinical features from unstructured text extremely reduced the total clinical feature space by around 50%.

**Mapping Clinical Features to Genes**

Phenome network is constructed from gene to clinical features associations. As described in the previous step we normalized the clinical features to UMLS concepts, where each clinical feature has associated OMIM id. Further association of genes to features is done through OMIM id from ‘mim2gene’ (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/ mim2gene) dataset.

**Figure 13**: Portion of Bio-RDF for gene DMD based on the DCO ontology. The upper network is the ontology providing the required semantics for the lower RDF network consisting instance data.

**Generating RDF**

The Resource Description Framework (RDF), an official W3C recommendation, provides a generic framework based on directed acyclic graph (DAG) to describe web resources. It is a semi-structured data model in which complex relations can be readily modeled [37]. RDF statements describe a resource, the resource’s properties and the values of those properties. Each statement is referred to as a “triple” that consists of a subject, predicate (property), and object (property value). Statements in RDF can be represented as graph of nodes (resources) connected by edges (properties) to values. For example the triplet, `<‘ATM’ ‘is a’ ‘Gene’>`, expressing...
‘ATM’ as subject, ‘is a’ the property and ‘Gene’ as object of the statement. Disease Card Ontology (DCO), an ontology currently under internal development to model and help relate mechanisms of actions (pathways) to biological entities, influence of genotypes and clinical findings that are operative in a diseased state is used to provide the required semantic framework in generating RDF. DCO is being developed using Protégé [110] in OWL, a language layered on top of RDF to offer support for axioms and inference. Jena (www.jena.sourceforge.net), a java framework for building Semantic Web applications is used to generate the required triples for RDF.

In the current version the data is retrieved directly from local relational databases to create RDF dynamically on the fly for the specific disease and gene set under study. However, the future versions will access a native RDF triple store to extract large subsets of graphs for a particular disease and gene set. The data includes genomic information (pathways and gene ontology annotations) and phenomic information (OMIM, Syndrome DB clinical features and Mouse Phenotypes) associated with the test genes under study (Figure 10). Figure 13 provides a portion of DCO and associated RDF. As here we are focusing on CVD, the mouse phenotypes are restricted under ‘cardiovascular system phenotype’ parent term from the Mouse Phenotype Ontology.

**Ranking on Semantic Web (SW)**

Discovering relevant knowledge and developing effective information retrieval techniques are crucial towards realizing the vision of Semantic Web. Our ranking approach is based on an earlier work which was successfully implemented in BioPatentMiner System [104]. The same logic can be applied in finding the disease genes from an integrative functional Bio-RDF
network. In the next few sections, a brief overview is provided in considering the metrics for ranking resources. For a more complete in-depth analysis and formulae for the algorithm, refer to the original paper [98].

**Resource Ranking Importance**

Resource importance, scoring network elements according to their importance within the network structure, can be calculated by relationships it has with other resources on the SW. It explains that the meaning of many other resources have to be defined with respect to this resource. In the context of SW, two important metrics have been defined to estimate the importance of each resource, subjectivity (SS) and objectivity scores (OS) parallel to Kleinberg’s [98] hub and authority scores for the WWW graph (Figure 14). Kleinberg not only considers in-degree and out-degree for each node but also the importance of linked nodes. Accordingly, if a node is pointed to by a node with high SS, its OS increases. Similarly, if a node points a node with high OS, its SS is increased. Nodes with high subjectivity/objectivity scores are subject/object of many RDF triples.

**Significance of Subjectivity (SW) and Objectivity Weights (OW)**

In the present WWW, all links are of equal weight and considered equally important while calculating hub and authoritative scores. But the SW space is more complex, where each property might not be equally important and depends on the subject and object it is associated with. For example, consider the property in_pathway where it links a gene to a pathway it has role in. A gene associated with multiple pathways is more important compared to a pathway having many genes because a modification of the multipathway-
linked gene could affect many different pathways. Figure 15 illustrates the significance of semantic weights on gene-pathway association. On the other hand, the property associated_process links a mouse phenotype to a biological process.

As in the previous example, a biological process associated with multiple phenotypes is more important compared to a mouse phenotype having multiple processes. Therefore each property in SW space has pre-defined subjectivity and objectivity weights, which control the subject/object scores (resource importance) of the property. From the above examples, properties like in_pathway have higher subjectivity weight and properties like associated_process have higher objectivity weight. In our case gene is the subject for all the triples and each property is assigned a subjectivity weight (SW) of 0.9 and objectivity
Figure 15: a) Illustrating the significance of a gene associated with multiple pathways is considered more important compared to a pathway having multiple genes b) Assigning subjectivity and objectivity weights to the property ‘associatedPathway’ for the triple ‘gene – associatedPathway – Pathway’ weight (OW) of 0.1. The assumption is that sum of SW and OW must be equal to 1. For a more comprehensive description of the algorithm, refer to the original paper [98].

**Ranking the Retrieved Results**

Search result ranking is an important research area in Information Retrieval. The results are not determined by specific query but by the importance of the results on the overall information space. We used ARQ (http://jena.sourceforge.net/ARQ/), a query engine for Jena that supports SPARQL RDF query language. However, SPARQL doesn’t prioritize the results, we borrowed a technique from M. Sougata et al [40, 104] which adds an extra computational layer to rank the
results. For each query the SPARQL returns a set of variable bindings matching to the query parameters and each unique result produces a graph formed from the triples matching the criteria. We extract the associated graph and compute a score for every result.

**Results**

**Benchmark of the method**

To explore the feasibility of our approach in candidate gene prioritization, we randomly selected 60 diseases from a total of 423 CVD from OMIM database having known gene relations and associated clinical synopsis. The algorithm was not provided with any obvious link between target gene and the disease as we want to make sure that our method detects the true functional relationship between the disease and the gene. For each disease, we pulled out the genes located in the locus specified in the OMIM. On average each region contains around 300 genes. The benchmark results were promising, as for 44 out of 60 cases (74%) the related gene is ranked in the top 10 and in 33 cases (55%) ranked in top 5.

**Application**

We tested the efficacy of our method in prioritizing candidate genes from CVS Disease – implicated genomic regions and differentially expressed genes from expression studies.

**Prioritizing Candidate Genes from CVS Disease – Implicated Genomic Regions**

Linkage analysis is a successful method to associate diseases with specific genomic regions. However, these regions are often large, containing hundreds of genes, which make experimental
methods employed to hunt the disease gene difficult and costly. We used our integrative genome – phenome based ranking approach to prioritize candidate genes from the CVS disease – implicated genomic regions. As test sets, we used known genes from 2 loci recently implicated in cardiomyopathy[105, 111].

Prioritization of Genes at a Locus for Hypertrophic Cardiomyopathy on Chromosome 7p12.1-7q21

We ranked the 110 genes occurring in the chromosome locus 7p12.1-7q21 (~27.2 megabases), a recently reported inherited cardiomyopathy susceptibility region on human chromosome 7 [105]. Mutations in the top ranked genes, namely, ELN, GTF2I, GTF2IRD1, BAZ1B and LIMK1 (in mouse or human or both) have been associated with Williams-Beuren Syndrome, a syndromic disease characterized by infantile hypercalcemia, supravalvar aortic stenosis (OMIM ID: 194050) and less frequently hypertrophic cardiomyopathy [112].

Prioritization of genes at a locus for Dilated Cardiomyopathy on chromosome 10q25-26

After the prioritizing the 68 genes in the chromosome 10q25-26 region (~9.5 megabases, locus for cardiomyopathy, diffuse myocardial fibrosis and sudden death)[30] the top ranked gene was FGFR2. FGF signaling via FGFR2 regulates myocardial proliferation during midgestation heart development and in the absence of this signal, newborn mice develop dilated cardiomyopathy[113]. From the study[114], comparing the GRK5 (ranked second) expression in patients with left ventricular volume overload disorders and dilated cardiomyopathic hearts, their exists a relation between the expression of GRK5 and alterations in myocardial beta-adrenoceptor signaling in volume overload. The result point to myocardial GRK5 regulation in
cardiac disease localized to ventricles. Jahns et al have provided direct evidence that an autoimmune attack directed against the cardiac beta(1)-adrenergic receptor, ADRB1 (ranked third) may play a causal role in DCM[115]. A recent study reports the use of ADRB1 as prognostic marker, risk predictor and clinical relevance of stimulating antibeta1-AR in DCM[116].

Prioritizing Candidate Genes from the Differentially Expressed genes in CVS Disease

Microarray analysis is a powerful technique for high-throughput, global transcriptomic profiling of gene expression. It holds great promise for analyzing the genetic and molecular bases of cardiovascular diseases and various other complex diseases and permits the analysis of thousands of genes simultaneously, both in diseased and non-diseased tissues and/or cell lines [117]. To test our prioritization approach, we used the dataset of differentially expressed genes in human idiopathic dilated cardiomyopathy [106].

Gene Prioritization of Differentially Expressed Genes in Human Idiopathic Dilated Cardiomyopathy (DCM)

We used our prioritization approach to rank 216 differentially expressed genes from the expression profiles of myocardiac biopsies from 10 DCM patients [106]. The top gene is DMD, which is well known in cardiac function and malformation. Specific DMD gene mutations may protect against or inhibit development of DCM. K336E mutation in ACTA1(Ranked 2) is associated with fatal hypertrophic cardiomyopathy[118]. A missense mutation of CRYAB (Ranked 5), Arg157His, was found in a familial DCM patient and the mutation affected the evolutionary conserved amino acid residue among alpha-crystallins[119]. Although GJA1
(ranked 8th) is not associated with hypertrophic cardiomyopathy, disturbances in Cx43 expression and localization are reported to influence heart embryogenesis and maturation and contribute to hypertrophy and dysfunction of the right ventricle, including arrhythmias in children with tetralogy of fallot[120]. RYR2, ranked 10th in our list, encodes ryanodine receptor found in cardiac muscle sarcoplasmic reticulum. Mice with the R176Q cardiac RYR2 mutation exhibit catecholamine-induced ventricular tachycardia and cardiomyopathy [121]. RYR2 mutations are also known to cause cardiomyopathies and sudden cardiac death [122].

**Advantages of Using Semantic Web Technologies**

*Flexible Integration of Genome to Phenome Networks and Querying*

An important feature of our framework is the ability to include multiple data sources related to different disease features for prioritization. Here RDF, being a DAG, provided us a very flexible way to integrate different layers of information and also to mine the integrated network by applying graph theoretical algorithms. We accessed whether our sequential integrative genomic – phenomic approach is capable enough in prioritizing the
**Figure 16:** Rank ROC curves for validation of integration - Each curve is generated based on rankings of the implicated gene out of 300 genes (on average) from the loci associated with 60 sample OMIM diseases. The 4 curves, as indicated by different colors are associated with sequential integration of different genome – phenome data sources. The data sources used to construct every ROC curve are indicated on the figure.

Implicated gene for the 60 sample diseases. Sensitivity and specificity values were computed for each of the 60 prioritizations using the methodology described by Aerts et al[32]. Sensitivity refers to the frequency (% of all prioritizations) of disease implicated genes that are ranked above a particular threshold position. Specificity refers to the percentage of genes ranked below this threshold. We plotted rank receiver operating characteristic (ROC) curves to prove that increasing the number of heterogeneous data sources enhances the probability in predicting the disease implicated gene. ROC curves from Figure 16 illustrate that sequential addition of genome – phenome data integration improves the overall performance of ranking. The bigger the area under the curve (AUC) the better will be the performance and from Figure 16, the area under the curve associated with all the data sources is comparatively bigger than other curves thus supporting our hypothesis.
In addition to ROC curves, the following specific example explains how RDF based integrative approaches aided us to home in on the gene SDHB underlying Paragangliomas 4 (OMIM ID: 115310). This disorder has several cardiovascular system symptoms (palpitations, tachycardia, hypertension) allowing us to include it in the list of CVD. SDHB is one of the 245 genes located at the genomic region 1p36.1-p35. Figure 17 illustrates how flexible integration provided by RDF improves the rank of implicated gene. As a general conclusion, using RDF facilitates to add in a flexible and modular manner additional disease – specific data sources to enhance its overall performance. Moreover, the algorithm also requires constant traversals of graph to score each node in the network and SPARQL provides the required graph querying capabilities.

**Figure 17:** Sequential addition of Genome – Phenome datasets improves SDBH gene ranking implicated in PARAGANGLIOMAS 4

**Adding Context through Semantic Weights**

As discussed in the methods section and from Figure-15, by incorporating context specific subjectivity (SW) and objectivity weights (OW), we were able to improve ranking of certain genes. We generated ROC curves as explained before with and without semantic weights (Figure 18 a) by including all data sources and with an average of 300 gene set per disease. Tests were
also performed to check the overall performance variation (Figure 18b) with respect different combination of subjective and objective weights. As mentioned before keeping more subjective weight (for specific properties) increased the performance of final ranking. It’s clearly visible that adding appropriate semantic weights, as explained in the methods section, improved the overall performance in ranking. For example, the ACADVL gene implicated in mitochondrial very-long-chain acyl-CoA dehydrogenase deficiency (OMIM ID: 201475) ranked 53 without any Subjectivity and Objectivity weights, but improved its ranking to 9 after adding weight functions.

**Figure 18:** Rank ROC curves for validation a) each curve is generated based on rankings of the implicated gene out of 300 genes (on average) from the loci associated with of 60 sample OMIM diseases including all data sources. The curves provide the comparison of overall ranking performance with and without semantic weights b) each curve is generated based on different combination of weights. The curve provides evidence of importance in giving more subjective weight for particular properties

*Ability to Investigate Other Resources (apart from genes) in Bio-RDF*

As all resources in the integrated Bio-RDF information space are ranked, we can issue further SPARQL queries to retrieve any ranked list of resources. Using the Human Idiopathic DCM
example, we investigated further by querying for relative pathway ranking. This provides further evidence of other important entities shared in the network to corroborate our initial findings. Figure 19 illustrates each SPARQL query and pathways returned from multiple sources. This feature is particularly useful for expression studies as the differentially expressed genes are already related in a particular disease context.

**Discussion**

Our approach to enrich lists of gene or candidate gene prioritization differs from other methods in multiple ways, right from coverage of data sources, data integration methods and applied mining algorithms. To the best of our knowledge, apart from G2D [33] and PROSPECTR[95] and POCUS [96], most of the current tools to enrich lists of genes or candidate gene prioritization use training gene set. But in many cases, training gene sets are not available and results are highly dependent on the quality of training set used. G2D uses MeSh (www.nlm.nih.gov/mesh) disease terms, from publications associated with each OMIM disease, as disease clinical features. These features are not comprehensive or granular compared to the clinical synopsis section we used, limiting the potential of G2D. In addition, none of the current approaches integrate human and mouse clinical features although the mouse is the key model organism for the analysis of mammalian developmental, physiological, and disease processes [123]. Our methodology has two phases, first to find the biologically functional important genes from the test set by
integrating multiple genomic data sets. The importance is scored based on their participation in multiple pathways, biological processes and molecular functions independent of any particular disease. Next, we include specific disease context to the genomic network by adding phenotypic or clinical features relevant to the disease under study (Ex: All Cardiovascular symptoms associated with the test genes from OMIM). This step increases the ranking of those specific genes, considered important from earlier case and also associated with clinical features related to the disease under study. In general, we are applying network centrality analysis to rank resources according to their importance within the Bio-RDF network structure. Moreover, in this case, the importance of a resource is calculated by diverse data sets (from genome to phenome) integrated
into the information space. Additionally, resource ranking is performed semantically by including contextual semantic weights on the properties connecting the resources. Our approach however has some limitations. First, the prioritization can only be accurate as the underlying online sources from which the annotations are retrieved. Second, prioritization can be applied only on diseases where clinical features are available.

**Conclusion**

We have used for the first time in human disease gene prioritization combination of mouse phenotype and human disease clinical features from OMIM clinical synopsis. Apart from coverage of data sets used, we have shown how we can leverage on Semantic Web standards and techniques to apply on specific biological problem, right from RDF and OWL for integration, application of customized network centrality analysis algorithms for mining Bio-RDF and also retrieving ranked results using graph query languages such as SPARQL. Although, in the current study we focused on the cardiovascular system, our approach can be applied to any group of genes or disease sets. One immediate application could be in applying to OMIM diseases (around 1554) having known loci but unknown molecular basis. As the functional annotations of human and mouse genes improve we envisage a proportional increase in the performance of this approach. Finally, we strongly believe that our methods will accelerate the disease gene discovery process by gathering and sifting through all knowledge of each candidate gene including its homologs and their phenotype. This in turn will enable targeted research on how mutations in the gene contribute to disease and provide specific leads towards novel diagnostic and therapeutic approaches.
Chapter 7: Construction of Genome – Phenome Network

The fundamental aspect of this project is to efficiently overlay a similarity-based phenome network over a genome network which requires a common formalism to support efficient integration and knowledge mining of this network. RDF, a Semantic Web standard, provides a flexible yet efficient protocol to support these aspects. But, for any RDF generation requires template ontology to provide the required semantics to tag each instance present in RDF. Apart from data conversion, and applying page rank algorithm [101] onto Phenome – Genome network, we did case study analysis to verify the validity of our approach through literature and significant statistical analysis. This chapter elaborates further in precisely defining the logical problems in reaching our goals by encompassing the following topics:

- Construction of Disease card Ontology
- RDF models initiated to create a Genome and Phenome network
- Semantic Web databases and querying tools
- Description, conversion and loading phenome – genome data into Oracle 10g RDF store
- Extending ranking algorithm onto Genome – Phenome Network
- Information Flow and Framework

Construction of Disease Card Ontology (DCO)

Domain ontologies enable querying of heterogenous datasets, but they are not sufficient for constructing a phenome – genome framework as the data of interest commonly spans multiple domains. In order to provide an integrated platform for our framework to discovery new knowledge through seamless integration of the very diverse genomic, phenomic and pharmacome domains, we used an existing ontology under development. Disease Card Ontology
(DCO), a pilot ontology [124] panning phenome, genome and pharmacome networks to provide the required classes to tag the instances and properties semantically associated with various instances in our framework. A manual construction approach was adapted here due to the broad coverage and complexity of knowledge domain involved. Ontology editor Protégé [110, 125] is used as the primary tool for implementing OWL framework. To enhance the editing and visualization flexibility, several plug-ins were also used including PROMPTViz [126] for ontology comparison and merging and OwlViz[127] for visualization.

We have taken into account that the key factor in ontology development is to reuse knowledge components whenever possible. Therefore, previously existing ontology sources were thoroughly examined to select relevant reusable knowledge resources to allow efficient knowledge mapping and sharing among independent data sources. The key top-level knowledge components from these sources were examined and mapped to our knowledge framework. Therefore, data from these sources is naturally compatible to the DCO for integration to our knowledge base. High-order hierarchy of the DCO is attempted to be maintained to allow seamless data integration. We manually pruned irrelevant or duplicate branches and added new concepts/relationships to accommodate data needs. On the other hand, some of the concepts mapped from UMLS Semantic Network are not relevant for our study, such as the concepts pertaining to Manufactured Object (i.e. Medical Device) and Activity (i.e. Daily or Recreational Activity, Machine Activity). We removed the node for these objects but kept the children nodes for Clinical Drug (sub-node for Manufactured Object) and Occupational Activity (sub-node for Activity). The current DCO contains 649 classes and subclasses, with average sibling number of 11 per class. There are total of 175 properties, with 53 properties domain-specified, 46 range-specified, and 29 inverse-specified. Figure 20 presents a top-level view of the ontology concepts.
used for data integration in this study. Though the DCO is much comprehensive covering various domains, here we restricted our classes and properties required for phenome–genome framework.

**Figure 20:** High-order View of Disease Card Ontology[124]. A simplified schematic representation of ontology classes used for integrating biological, pharmacological, and phenomical data

**Generating RDF**

Disease Card Ontology (DCO) described earlier helps to model and provides the required semantic framework in generating RDF. Each sub element in RDF triple is identified with a Unique Resource Identifier (URI). Two resources are said to be identical if they have the same URI. This unique attribute simplifies data integration in RDF, as all the data associated between two identical resources is automatically merged. This structure is ideal enabling easy integration
for highly interconnected domains subjected to perpetual evolution and transformation. We
assigned our own custom name space URI ‘http://www.chmcc.org/DiseaseCards.owl#’ to
uniquely identify each resource in our –omics information space. As the first step to build RDF,
we created models based on the logic and semantic relationships defined in the DCO for each of
our phenome and genome domains (Figure 21). These models provide the required node and
edge relational mapping mechanism from the instance data to DCO and semantically annotate to
relate different –omic entities. From example, a disease from OMIM and its susceptibility
gene(s) can be connected by the property of “hasAssociatedGene”. Jena[128], a java
environment frame work for building Semantic Web applications is used to generate the required
triples in RDF/XML format. The following examples (Figure 22) explain how we differentiate
same id’s belonging to different entities and also how we differentiate pathways.
Figure 21: RDF data models (a) Genome RDF model (b) Phenome RDF model
Figure 22: RDF resource representation differentiating multiple entities

Converting phenome data to RDF:

The data and methods required for phenome map generation are explained in Chapter-5 and requires two types of data,

1. OMIM phenotype data

2. Phenomap matrix obtained from phenotype similarity measures

OMIM phenotype data

Each OMIM phenotype record encapsulates clinical features (from TX and CS section) and implicated genes where available. In our case (only CVS) the clinical feature CUIs obtained after semantic normalization are converted to RDF utilizing the classes and semantic relations from DCO. The abstract model depicting the entities and relations of the OMIM disorder phenome model is illustrated in Figure 21(b). This data mainly provides the required annotations to the phenome network obtained from phenomap.
Phenomap matrix to RDF:

Constructing a phenome network is not straightforward as the ‘phenomap’ has to be converted to RDF. Phenomap is the distance matrix obtained (From Chapter-5) after computing the similarities between clinical feature vectors by finding the cosine of angle between each pair of them. The basic element of RDF is a triplet, which supports only binary relations. It links two individuals or an individual and a value. Here, each triplet has to define the similarity between two phenotypes and also delineate the similarity score for such relation. The conundrum is how to represent properties of a property connecting subject and object of a triplet. This is a data modeling problem in RDF and a common solution is to use ‘n-ary’ relations (http://www.w3.org/TR/2004/WD-swbp-n-aryRelations-20040721/). This method creates an individual which stands for an instance of the relation and relates the things that are involved in that instance of the relation. We can think of the original relation then as a class of all these relation instances. Figure 23 illustrates the RDF model used to hold Type1-phenomap (similarity matrix of phenotypes having known gene etiology). We created a special class ‘PhenotypeSimilarityRelation’ in DCO to represent relations between two phenotypes. The individuals which are instances of this relation class relate each pair of phenotypes by their similarity score and type of phenomap matrix (Type 1 or Type 2). Using DCO and implementing ‘n-ary’ relations, Jena was used to convert Phenomap to RDF/XML format. Appendix B provides an example raw RDF/XML encompassing both phenomap and phenome annotation data.
Converting genome data to RDF:

The required data sources and methods to construct a genome RDF dynamically from relational data bases are explained in Chapter-6. But in order to efficiently integrate and construct phenome–genome framework, even genomic data has to be converted to RDF. Similar to phenomic data conversion, using DCO as a template, a RDF/XML version of genomic data is obtained. The abstract genomic RDF model is illustrated in Figure 21(a). Appendix B provides an example of genome data in pure RDF/XML format.

Semantic Web database and querying tools

To provide support for storage and querying RDF/OWL, several database systems have been developed. Few are open source RDF database systems, such as, Sesame (a.k.a OpenRDF)[55] and Kowari [56] while the Oracle RDF data model[129-131] is a feature of the Oracle Database and therefore a commercial product. Most of these database systems implement their version of RDF query languages in compliance with standard SPARQL specifications [132]. After thorough
analysis, we opted for commercial Oracle 10g release 2 RDF data model (http://www.oracle.com/technology/tech/semantic _technologies/ index.html) due to its native support for RDF graph model, support for persistence management, scalability, indexing of the RDF triples and availability of a query language for the RDF graph.

*Oracle RDF Data Model*

A new data model has been developed for storing RDF and OWL data in Oracle Database 10g Release 2 for an open, scalable, secure management platform support within their existing Relational Database Management System (RDBMS)[133]. This functionality is a further enhancement on recent Oracle Spatial Network Data Model (NDM), which is an Oracle solution to manage graphs where RDF triples are persisted, indexed and queried, in a similar manner to other object-relational data types. As it stores data in triplets, Oracle10g falls into the category of ‘*native RDF repository*’. The RDF database supports three types of database objects: model (RDF graph consisting of a set of triples), rulebase (set of rules), and rule index (entailed RDF graph)[130, 134]. All uploaded RDF triples are parsed and stored as one universe in Oracle RDF database. The triples are stored in the system as entries in tables under MDSYS schema. Each RDF triple (subject, predicate, and object) is treated as one database object. Therefore, each RDF document containing multiple triples will result in many database objects. Every subject and object are mapped to nodes in the network, and predicates are mapped to network links that have their start node and end node as subject and object, respectively.
Oracle – SPARQL queries

In Oracle 10g RDF store, SDO_RDF_MATCH table function provides the required graphical query capabilities to search for an arbitrary pattern against the RDF data, including inferencing, based on RDF, RDFS and user defined rules. This function allows a graph query to be embedded in a SQL query and has been designed to meet most of the requirements identified by W3C in SPARQL [132] for graph querying. The query can be understood as finding a path in the RDF graph using a predetermined set of semantic relationships. The anatomy of SDO_RDF_MATCH table function is provided in Figure 24 and has the following attributes

\[
\text{SDO_RDF_MATCH (}
\begin{align*}
\text{query} & \text{ VARCHAR2,} \\
\text{models} & \text{ SDO_RDF_MODELS,} \\
\text{rulebases} & \text{ SDO_RDF_RULEBASES,} \\
\text{aliases} & \text{ SDO_RDF_ALIAS,} \\
\text{filter} & \text{ VARCHAR2} \\
\end{align*}
\) \text{ RETURN ANYDATASET;}
\]

**Figure 24:** Anatomy of ORACLE – SPARQL query[131]

The ‘query’ attribute is a string literal with one or more triple patterns, usually containing variables. A triple pattern is a triple of atoms enclosed in parentheses. Each atom can be a variable, a qualified name that is expanded based on the default namespace and the value of the alias parameter, or a full URI. In addition, the third atom can be a numeric literal, a plain literal, a language-tagged plain literal, or a typed literal. The ‘models’ attribute identifies the RDF model or models to use. The ‘rulebases’ attribute identifies one or more rule bases whose rules are to be applied to the query. ‘The models’ and ‘rulebases’ together constitute the RDF data to be queried. The ‘alias’ attribute identifies one or more name spaces, in addition to the default
namespaces, to be used for expansion of qualified names in the query pattern. The ‘filter’ attribute identifies any additional selection criteria.

**Loading and formation of phenome – genome network in Oracle RDF Store**

Oracle RDF loaders (http://www.oracle.com/technology/tech/semantic_technologies/sample_code/files/sdordf_converter.zip) support RDF only in N-triple format, thus both the genomic and phenomic data are to be converted into N-triple format using Jena API and further loaded into database. Appendix A provides the required RDF graphical models for each type of genomic and phenomic data source with their associated model and table name used to store the N-triple data. These RDF models form the basis to generate RDF/XML or N-triple formatted data. As part of phenomic data, approximately 1 million triples were generated by converting Type1- phenomap into N-triple format. Though we uploaded phenome and genome data independently both can be merged due to unique resource identification possible through Uniform Resource Identifier (URI) (www.w3.org/Addressing/) for each resource, any two resources having the same URI are assumed to be the same entity and can be merged. Here, gene Id’s are the key entities that are associated with both phenome and genome domain, thus forming a bridge between these two datasets. As, gene Ids are assigned the same URI in both the domains, once uploaded, both the phenome and genome RDF graphs are automatically connected together through common genes, thus forming a massive directed labeled graph instantiating ‘phenome – genome’ network inside the Oracle RDF store. Figure 25 illustrates the real model workflow of loading and forming Phenome – Genome network in the Oracle RDF store.
Figure 25: Real work flow model of creating, loading and forming Phenome – Genome network in the Oracle RDF store (Seamark Demonstration Image)

Extending the SW ranking algorithm onto the genome – phenome network

In Chapter – 6, we implemented centrality based extended page-rank [99, 101] related ranking algorithm on RDF to test its efficacy in prioritizing known implicated genes from loci associated with OMIM diseases (only 455 OMIM CVS). After thorough testing and verification of various parameters in the algorithm, we extended it onto the phenome – genome network. The ranking algorithm cannot be applied onto the complete integrated network in the Oracle RDF information space. Therefore, we extract phenome - genome sub-network (computable network) associated
with the clinically similar phenotypes of the query OMIM syndrome using Oracle SPARQL (SDO_RDF_MATCH function) queries. The sub-phenome network (Figure 26) consists of clinical features associated with disease group and their implicated genes. The computable sub-genome network (Figure 27) consists of primary and secondary genes with their associated pathways and gene ontology entities. The computable integrated phenome – genome is illustrated in Figure 28. The entire technique as illustrated in Figure 29 can be summarized as two step process in ranking relevant biological entities related to the phenome cluster under study.

- Construction of an in-memory model of a sub phenome - genome network
- Applying ranking algorithm onto this in-memory model

The integrative network in Oracle RDF store can be queried for any OMIM CVS and its related disorders starting from phenome network. This sub-group of queried diseases forms the phenome cluster understudy and the associated sub phenome - genome network can be extracted from the complete network in Oracle RDF store. Creation of sub genome network happens into two levels (Figure 30). The first level is constructed from the primary genes which form the seeds to this genome sub-network. Primary genes are those that are directly implicated with each phenotype in the phenome cluster. All the associated annotations for these primary genes such as interacting partners, pathways and gene ontology entities are further queried (from Oracle RDF) and are wrapped up in sub- network. In the second level, interacting partners to the primary genes, the secondary genes, form the seed genes and the congruent genome annotations are retrieved back from Oracle RDF and added to level-1.
**Figure 26**: Computable sub-phenome network

**Figure 27**: Computable sub-Genome network
All the pathways from different sources (from KEGG, Biocarta and Reactome) and all three different categorical types of gene ontology terms (biological process, molecular function and cellular components) constitute the annotations for the primary and secondary genes. Though sub phenome and genome networks are created independently, they are automatically connected in the stored in-memory Jena ontology model with their common genes (Figure 25). The ranking algorithm [98] is applied onto this sub ontology model to prioritize relevant biological entities which can be individual genes or functional modules such as pathways and gene ontology entities. All the entities in the sub-network are ranked based on their type of relations and also the number of relations with the other entities in the network.

**Figure 28:** Computable sub phenome – genome network
A much more fine explanation of the parameters and implementation of this ranking algorithm can be found in Chapter –6. Finally, SPARQL queries are issued onto the in-memory model using Jena – ARQ SPARQL engine[135] to retrieve and rank the associated entities based on their scores. Query results are ranked by calculating the specific score of each sub-network associated with each result. This can be achieved by using ‘CONSTRUCT’ query form in SPARQL which returns a single RDF graph (sub-network) based on the matching graph template specified in the query for each query solution in the solution sequence by substituting for the variables in the graph template, and combining the triples into a single RDF graph by set union. Figure 29 explains the work flow of the entire process applied in executing any particular OMIM CVS.
Figure 30: Two level genome network associated with each disease in the cluster
Chapter 8: Case studies and validation

Case studies were performed to test the efficacy of integration and implementation of ranking to discover and prioritize the following novel relationships present in the phenome – genome network.

- Disease – gene
- Disease – pathway
- Disease – Molecular Function
- Disease – Biological processes

In OMIM, the first category of 455 CVS phenotypes already have at least one implicated gene, where other modifier genes and underlying modules such as novel pathways, other molecular functions and processes can be dissected using our integrative framework by finding similar phenotypes to the query phenotype. Around 3500 OMIM CVS are of unknown etiology belong to second category, in which a subset (520 OMIM records) belongs to CVS. For this second category of syndromes of unknown etiology we can identify the other disorders sharing similar clinical features, a subset of which may be previously associated with human genes. Therefore, known causative genes from syndromes that are phenotypically similar to a genetically uncharacterized syndrome (second category) can be used to query the gene network to discover and prioritize functional modules. There is no database existing to provide curated relations between phenotype groups and associated biological entities to test our framework. So, we opted for case study analysis by pooling discovered established relations between phenome groups and associated modules from literature. Apart from literature validation of our computational predictions, several statistical tests were performed on ranks & scores obtained from the
permutation of the original genome RDF. The goal of permutation is to prove that the ranks and scores associated with the entities from real network (real ranks & real scores) are statistically significant. The next section explains permutation procedure and its significance.

**Semantic Permutation**

Scale-free networks have few vertices with higher degree and follow a power – law tail degree distribution (Figure 31) [136, 137]. Similarly, our sub-genome network have existence of few nodes with higher degree and follows a power law distribution (Figure 32), indicating scale-free properties that are inherent in other biological networks. In order to test the validity of our computed entity ranks and scores in the real genome network, we sought

![Log – Log plot showing Scale- free nature of networks](image)

**Figure 31:** Log – Log plot showing Scale- free nature of networks [137]
to permute it by keeping the degree distribution constant. In order to perform any statistical tests, the real and permuted graphs must have the same topology such as preserving scale free properties. But here the network is RDF, which is a semantic graph, heterogeneous in nature of having different types of property relations connecting different types of genomic entities (nodes). As explained earlier, our ranking approach determines score to any particular entity based on the different relations and also combination of relations it has with different entities, mimicking the true environment of a cell. Our important assumption is that the score or significance assigned to an entity must not be based on number of edges, but the number of each type of edges and the particular combination types of edges. Therefore, we semantically permuted the original graph as illustrated in Figure 33 and checked for its topology whether it’s conforming to the real network. As expected permuted graph is ‘scale-free’ considering each ‘typed degree’ distribution being identical as the real network.
Figure 33: RDF network permutation procedure. Total degree for each ‘gene’ node is decomposed into its ‘typed degree’ belonging to type of edges. ‘Typed degrees’ are permuted in such a way to keep the total ‘type degree’ distribution constant. Objects nodes (pathways, GO terms) are randomized before adding to their specific property.

Permuted networks did produce the required topological and structural properties featured in real networks, considering ‘typed degree’ distribution being identical as of real network and Figure 34 illustrates the scale – free nature of two separately permuted networks. ‘Typed degree’ can be defined as the total number of particular type of edges for a node, where a ‘degree’ of a node is the number of edges irrespective of their type for a node. As summary, in permuted RDF, we keep two parameters constant as of real networks. First, keep the “typed degree” constant and second, every node from the original RDF is present and also being connected to at least one other node.
**Figure 34:** Illustrating the scale – free nature of permuted networks (for Marfan Syndrome group)

<table>
<thead>
<tr>
<th>Genomic entity Id</th>
<th>Rank from real network</th>
<th>Ranks from Random network</th>
<th>Score from real network</th>
<th>Scores from Random network</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$R_1$</td>
<td>$R_{1,1}, \ldots, R_{1,1000}$</td>
<td>$S_1$</td>
<td>$S_{1,1}, \ldots, S_{1,1000}$</td>
</tr>
<tr>
<td>2</td>
<td>$R_2$</td>
<td>$R_{2,1}, \ldots, R_{2,1000}$</td>
<td>$S_2$</td>
<td>$S_{2,1}, \ldots, S_{2,1000}$</td>
</tr>
<tr>
<td>3</td>
<td>$R_3$</td>
<td>$R_{3,1}, \ldots, R_{3,1000}$</td>
<td>$S_3$</td>
<td>$S_{3,1}, \ldots, S_{3,1000}$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$n$</td>
<td>$R_n$</td>
<td>$R_{n,1}, \ldots, R_{n,1000}$</td>
<td>$S_n$</td>
<td>$S_{n,1}, \ldots, S_{n,1000}$</td>
</tr>
</tbody>
</table>

**Table 8-1:** Input data set for each test statistic after 1000 randomizations of real network for each type of entity
Where:

n: Genomic entity id, which can be a gene, pathway, biological process or molecular function

\( R_n \): Rank of genomic entity ‘n’ from real network

\( R_{n,1} \): Rank of genomic entity ‘n’ from 1\(^{st}\) randomization

\( R_{n,1000} \): Rank of genomic entity ‘n’ from 1000\(^{th}\) randomization

\( S_n \): Score of genomic entity ‘n’ from real network

\( S_{n,1} \): Score of genomic entity ‘n’ from 1\(^{st}\) randomization

\( S_{n,1000} \): Rank of genomic entity ‘n’ from 1000\(^{th}\) randomization

From Table 8-1, \( R_{1,1} \) \( \ldots \) \( R_{n,1} \) (Ranks from randomized networks) are permutation of real ranks \( R_I \) \( \ldots \) \( R_n \) (Ranks from real network) of 1\(^{st}\) round mutation (randomization). Here, we briefly explain different statistical tests and interpretations used in our statistical analysis.

**Statistical Tests**

*Significance of entity based on its score in real network:*

As we are randomizing the original network keeping the degree distribution of each type of edge (property relation) constant, we calculated P-value by comparing the 1000 randomized scores \( S_{n,1} \) \( \ldots \) \( S_{n,1000} \) with the original score \( S_n \) of the entity. Our objective here is to find the significance of each entity by its score in the real network irrespective of its rank. P-value defines the value relation between scores from real and random networks for each entity and is calculated using the below formula,

Significance of original score, P-value = \( X / \text{Number of randomizations} \)
Where,

\[ X = \text{Total number of random scores better than or equal to original score} \]

Number of randomizations = 1000

We consider the entity significant in the real network if its P-value < 0.05.

**Correlation test:**

Correlation test is performed to test the correlation between real ranks (ranks from real network) and random ranks (ranks from random network). The idea behind this test is to verify whether the real ranks correlate with random ranks. We expect our real ranks significant if they don’t correlate with the random ranks. Table 8-1 provides the input data format for such analysis. Let \( \Gamma \) be (rank) correlation between real ranks and ranks from mutated networks (randomized networks).

\[
H_0: \Gamma = 0 \\
H_1: \Gamma \neq 0
\]

**Interpretation:** \( H_0 \) is true signifies that the ranks conferred on the entities by the mutated procedure are chaotic and the ranks from real network are reliable. \( H_1 \) is true and \( \Gamma > 0 \) means that the methods used to rank from real network and randomized network are equally effective in ranking entities.

We test the validity of \( H_0 \) versus \( H_1 \) and let \( \gamma \) be sample correlation coefficient.

**Test statistic:**

\[
t = \frac{\gamma}{\sqrt{1-\gamma^2}} \times \sqrt{\text{df}}
\]

df: degrees of freedom = total entities – 2
Test:

Reject $H_o$ in favor of $H_1$ if,

$|t| > t_{0.05, df}$ is 5% critical value of $t$-dist with ‘total entities -2’ degrees of freedom

P-value = $Pr(t_{df} > |t|/H_o)$.

df: degrees of freedom = total entities – 2

After sample rank correlations are calculated, p-value are determined for each randomization (1…1000). If in some cases $H_o$ is rejected, we perform meta-analysis to associate a single P-value for entire experiment.

Meta analysis based on Uniform Distribution methods [138] :

$H_o$: $\Gamma = 0$

$H_1$: $\Gamma \neq 0$

Test statistic:

1000 randomizations produce 1000 P-values: $p_1$, $p_2$, $p_3$, $p_4$ ........ $p_{1000}$. Theoretically these P-values are uniformly distributed under $H_o$.

Let $P_{\min} = \min \{ p_1, p_2, p_3, p_4 ......... p_{1000} \}$

$P_{\text{test}} = 1 - \left(1 - v\right)^{\frac{1}{1000}}$

Where, $v = 0.05$, is the prescribed level.

Overall conclusion:

Reject $H_o$ if $P_{\min} < P_{\text{test}}$

If $H_o$ is accepted, it explains that in overall there is no correlation between real ranks and each set of ranks from 1000 randomizations.
Chi – Square method \( (x^2) \) method

This test is performed to verify the validity of real ranks by checking the distribution of random ranks (ranks from random networks). Possible values of random ranks from randomizations (1… 1000th) for nth entity (from Table –1): \( R_{n, 1} \ldots \ldots R_{n, 1000} \)

\[ H_{0i}: R_{n, i} \sim \text{Uniform distribution over } R_{n, 1} \ldots \ldots R_{n, 1000} \]

\[ H_{1i}: H_{0i} \text{ is not true} \]

Interpretation:

\( H_{0i} \) true means that \( R_{n,i} \) (random ranks) are uniformly distributed over 1000 randomizations. The uniform distribution is the most chaotic (with maximum entropy) distribution. By testing whether the random ranks are uniformly distributed, validity of real rank can be justified. If the random ranks are uniformly distributed or not hovering around the real rank, then ranks from real network are significant.

Data for testing \( H_{0i} \) versus \( H_{1i} \):

The real network is mutated 1000 times and the data required for analysis are \( R_{n,1} \ldots \ldots R_{n,1000} \) (from Table – 1) random ranks, where \( R_{n,i} \) are sample observations of sample n.

A frequency distribution of random ranks is prepared.

<table>
<thead>
<tr>
<th>Ranks:</th>
<th>1 – p</th>
<th>(p+1) – q</th>
<th>(q+1) – m</th>
<th>(m+1) – n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Frequency:</td>
<td>f1</td>
<td>f2</td>
<td>f3</td>
<td>f4</td>
</tr>
<tr>
<td>Expected Frequency:</td>
<td>e1</td>
<td>e2</td>
<td>e3</td>
<td>e4</td>
</tr>
</tbody>
</table>

Where,
n: total number of entities
p = bin or pool size
$f_i$: Frequency of ranks in that particular interval
$e_i$: Expected Frequency = (number of randomizations * p) / n

**Test Statistic:**

$$X^2 = \sum_{i=1}^{y} \frac{(f_i - e_i)^2}{e_i}$$  
Where: $y = $ total number of pools = total entities / p

**Test:**

Rejected $H_0$ if $X^2 > X^2_{0.05, df}$, Where $df = y - 1$

$P$-value = $Pr\left(X^2_{df} > X^2\right)$, Where $df = y - 1$

The test is carried for each genomic entity (Gene, Pathway, Biological process, molecular function) and $P$-values are computed for every entity.

**Bonferroni Correction:**

For entities associated with the phenome group understudy, if the $P$-value is not significant enough to accept null hypothesis, we calculate “Bonferroni Correction (BC)” and compared their respective $P$-values with this new ‘BC’.

As we are testing simultaneously null hypothesis equal to total number of entities, Bonferroni Correction = 0.05 / Total number of entities

Significance is given to those entities whose respective $P$-values are lesser than ‘BC’
**Z – Score:**

The purpose of this test is to check whether each genomic entity’s real network score is in proximity to the average of random scores. If the real score is in adjacent to the average of the random scores, it clearly undermines the importance of real score. Let \( (S_n) \) be the score computed from real network associated with the entity ‘n’. After 1000 network randomizations, let \((S_{n, 1} \ldots \ldots S_{n, 1000})\) be the 1000 scores associated with the entity ‘n’. The average \((X_{\text{ave}})\) of these 1000 scores is computed.

\[ H_0: X_{\text{ave}} = S_n \]

\[ H_1: X_{\text{ave}} > S_n \text{ or } X_{\text{ave}} < S_n \]

**Test Statistic:**

Reject \( H_0 \) at level \( \alpha \), if

\[
\left( \frac{X_{\text{ave}} - S_n}{SD} \times \sqrt{R} \right) > Z_{\alpha}, \text{ at } \alpha = 0.05, 1.645 \text{ (upper 100 x } \alpha \text{ percentile)}
\]

Where,

\( SD: \text{ Standard deviation of random ranks } (S_{n, 1} \ldots \ldots S_{n, 1000}) \)

\( R: \text{ Total randomizations (1000)} \)

**Test Statistic:**

\[
Z = \left( \frac{\hat{U} - R}{S} \right) \times \sqrt{TR}
\]

Where,

\( \hat{U} = \text{ mean of Random Scores for that particular entity} \)

\( S = \text{ Standard Deviation of random Scores for that particular entity} \)
R = Real Score

TR = total number of randomizations

Test:

Accept $H_0$, if $Z > Z_{0.05}$

P-value = $2 \times (1 - Z_{0.05})$

The test is carried for each genomic entity (Gene, Pathway, Biological process, molecular function) and P-values are computed for every entity.

Case Studies

To demonstrate our framework and potential of our methods, we conducted case studies on two different phenome groups,

- Marfan Syndrome group
- Obesity Disease group

Methods relevant to forming these clusters are explained in Chapter –5. The objective of our case study analysis is to computationally predict proven links, from literature, between these phenome groups to their functional modules and genes. The results are also presented with robust statistical validations explained in the previous section.

Case Study 1: Marfan syndrome group

Marfan’s syndrome is a clinically and heterogenous group of disorders and classified as a connective-tissue disorder, a type of disorder that may affect any one of the numerous steps in the biosynthesis and the metabolism of (connective tissue) or the processes by which the macromolecules are physically organized and oriented to one another[139]. Common traits include, but not limited to cardiovascular abnormalities (aortic regurgitation, mitral valve
prolapses, aortic root dilatation and congestive heart failure), ophthalmologic abnormalities (myopia, retinal detachment and early glaucoma) and skeletal abnormalities (kyphoscoliosis, spondylolisthesis and arachnodactyly). FBN1, fibrillin-1 gene is an extracellular-matrix protein and harbors missense mutations in two unrelated Marfan syndrome patients [140]. The other related disorders with overlapping clinical features include Loeys–Dietz syndrome [141, 142], Ehlers–Danlos syndrome[142], Shprintzen-Goldberg syndrome [4, 143], Ectopic lentis and Aortic aneurysm [143]. Apart from FBN1, mutations in transforming growth factor, beta 1 (TGF-β), transforming growth factor, beta receptor I, TGFBR1 and transforming growth factor, beta receptor II, TGFBR2 affect. Loeys et al [141, 142] provided direct evidence of abnormal signaling in transforming growth factors beta (TGF-beta) pathway in the pathogenesis of Marfan syndrome (MFS).[144]. Due to mutations in fibrilin and latent transforming growth factor β (TGF-β) binding proteins (LTBPs) extracellular matrix is affected[3]. To test whether our framework can successfully reproduce the facts curated from the literature (Figure 35), we issued a query into our phenome network to retrieve Marfan syndrome [OMIM: 154700] with its related disorders and further constructed a genome network to rank important biological entities ranging from genes, pathways and gene ontology terms associated with phenome group. Figure 36 provides the ranked retrieved results for each type of biological entity and Figure 37 illustrates the real RDF graph depicting the ranked entities. The graph is generated using Welkin RDF visualizer [145]. It can be observed that there is a 100% match of our computationally predicted results with the literature. The following section explains the statistical evaluation results for relevant validation.
**Figure 35**: Information about Marfan phenome group curated from literature

**Statistical Validation:**

We permuted the genome network associated with Marfan Syndrome group for 1000 times and statistical tests were performed to verify the significance of the biological entities curated from literature (genes, pathways, biological processes and molecular functions). Table 8-2 provides the total number of each type of entity present in this genome network.

**Correlation test:**

Sample correlations are calculated between each set of real ranks and random ranks per each randomization both for genes and pathways (1000 randomizations).
From Phenome - Genome Integrative Framework

**Disease Group**
- ACHONDROPLASIA, ACH
- AORTIC ANEURYSM, FAMILIAL THORACIC 4
- ARTERIAL TORTUOSITY SYNDROME, ATS
- EHLERS-DANLOS SYNDROME, TYPE I
- EHLERS-DANLOS SYNDROME, TYPE II
- EHLERS-DANLOS SYNDROME, TYPE III
- EHLERS-DANLOS SYNDROME, TYPE IV, AUTOSOMAL DOMINANT
- EHLERS-DANLOS SYNDROME, TYPE VI
- LARSEN SYNDROME, AUTOSOMAL DOMINANT; LRS1
- LOEYS-DIETZ SYNDROME; LDS
- MARFAN SYNDROME, TYPE II; MFS2
- MARFAN SYNDROME; MFS
- OSTEOGENESIS IMPERFECTA, TYPE I
- OSTEOGENESIS IMPERFECTA, TYPE IIA
- POLYCYSTIC KIDNEYS
- SHPRINTZEN-GOLDBERG CRANIOSYNOSTOSIS SYNDROME
- ECTOPIA LENTIS, ISOLATED
- SPONDYLOEPIPHYSEAL DYSPLASIA CONGENITA; SEDC

**Genes**

<table>
<thead>
<tr>
<th>ID</th>
<th>NAME</th>
<th>RANK</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>7046</td>
<td>TGFBR1</td>
<td>2</td>
<td>193.8945</td>
</tr>
<tr>
<td>7040</td>
<td>TGFBR1</td>
<td>3</td>
<td>160.5482</td>
</tr>
<tr>
<td>7048</td>
<td>TGFBR2</td>
<td>4</td>
<td>141.6868</td>
</tr>
<tr>
<td>2022</td>
<td>EN3</td>
<td>5</td>
<td>126.322</td>
</tr>
<tr>
<td>1277</td>
<td>COL1A1</td>
<td>6</td>
<td>116.222</td>
</tr>
<tr>
<td>5781</td>
<td>PTPN11</td>
<td>7</td>
<td>119.6009</td>
</tr>
<tr>
<td>1280</td>
<td>COL2A1</td>
<td>8</td>
<td>89.88508</td>
</tr>
<tr>
<td>2006</td>
<td>ELN</td>
<td>10</td>
<td>66.58788</td>
</tr>
</tbody>
</table>

**KEGG Pathways**

<table>
<thead>
<tr>
<th>ID</th>
<th>NAME</th>
<th>RANK</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04510</td>
<td>Focal adhesion</td>
<td>1</td>
<td>10.31677</td>
</tr>
<tr>
<td>hsa04010</td>
<td>MAPK signaling pathway</td>
<td>2</td>
<td>5.300823</td>
</tr>
<tr>
<td>hsa04512</td>
<td>ECM-receptor interaction</td>
<td>3</td>
<td>4.542034</td>
</tr>
<tr>
<td>hsa04350</td>
<td>TGF-beta signaling pathway</td>
<td>4</td>
<td>4.33663</td>
</tr>
<tr>
<td>hsa01430</td>
<td>Cell Communication</td>
<td>5</td>
<td>3.026584</td>
</tr>
<tr>
<td>hsa04810</td>
<td>Regulation of actin cytoskeleton</td>
<td>6</td>
<td>2.027113</td>
</tr>
<tr>
<td>hsa04900</td>
<td>Cytokine-cytokine receptor</td>
<td>7</td>
<td>1.948139</td>
</tr>
<tr>
<td>hsa04520</td>
<td>Adherens junction</td>
<td>8</td>
<td>1.827349</td>
</tr>
<tr>
<td>hsa01410</td>
<td>Cell cycle</td>
<td>9</td>
<td>1.012446</td>
</tr>
<tr>
<td>hsa04670</td>
<td>Leukocyte transendothelial</td>
<td>10</td>
<td>0.941665</td>
</tr>
</tbody>
</table>

**GO - Mol.Function**

<table>
<thead>
<tr>
<th>ID</th>
<th>NAME</th>
<th>RANK</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO_0005515</td>
<td>protein binding</td>
<td>1</td>
<td>18.9638</td>
</tr>
<tr>
<td>GO_0005509</td>
<td>calcium ion binding</td>
<td>2</td>
<td>11.1334</td>
</tr>
<tr>
<td>GO_0005201</td>
<td>extracellular matrix struct</td>
<td>3</td>
<td>10.5289</td>
</tr>
</tbody>
</table>

**Biocarta Pathways**

<table>
<thead>
<tr>
<th>ID</th>
<th>NAME</th>
<th>RANK</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>h_albPathway</td>
<td>ALK in cardiac myocytes</td>
<td>1</td>
<td>1.65263</td>
</tr>
<tr>
<td>h_tgfbPathway</td>
<td>TGF beta signaling pathway</td>
<td>2</td>
<td>1.465765</td>
</tr>
<tr>
<td>h_mapkPathway</td>
<td>MAPK kinase Signaling Pathway</td>
<td>3</td>
<td>0.893513</td>
</tr>
<tr>
<td>h_p38mapkPathway</td>
<td>p38 MAPK Signaling Pathway</td>
<td>4</td>
<td>0.800154</td>
</tr>
<tr>
<td>h_runsPathway</td>
<td>NFkB activation by Nontypeable</td>
<td>5</td>
<td>0.748297</td>
</tr>
<tr>
<td>h_medPathway</td>
<td>Signaling of Hepatocyte Growth</td>
<td>6</td>
<td>0.582204</td>
</tr>
<tr>
<td>h_slrPathway</td>
<td>Function of SLRP in Bone</td>
<td>7</td>
<td>0.435121</td>
</tr>
<tr>
<td>h_to1Pathway</td>
<td>Role of Tob in T-cell activation</td>
<td>8</td>
<td>0.396068</td>
</tr>
<tr>
<td>h_integrinPathway</td>
<td>Integrin Signaling Pathway</td>
<td>9</td>
<td>0.380822</td>
</tr>
<tr>
<td>h_agrPathway</td>
<td>Agrin in Postsynaptic</td>
<td>10</td>
<td>0.346665</td>
</tr>
</tbody>
</table>

**Table 8-2: Entity statistics of genome**

BioRDF associated with Marfan Phenome

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Total Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>305</td>
</tr>
<tr>
<td>Pathway (BIOCARTA)</td>
<td>135</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>46</td>
</tr>
<tr>
<td>Pathway (REACTOME)</td>
<td>14</td>
</tr>
<tr>
<td>Biological Process (GO)</td>
<td>266</td>
</tr>
<tr>
<td>Molecular Function (GO)</td>
<td>175</td>
</tr>
</tbody>
</table>

Figure 36: Predicted from clustering and ranking associated genome-BioRDF of the group
**Figure 37:** RDF graph associated with Marfan syndrome group with visual depiction of ranked entities

**Correlation test for Genes:**

Figure 38 is the histogram of correlation P-value frequencies with bin size 0.05 between real and permuted ranks of genes and only 1.6% (16 correlation tests out of 1000) of times null hypothesis is rejected (P-values are less than 0.05). In order to summarize the overall correlation criteria between real and random ranks, we performed meta-analysis over the individual correlation values.

P_{\text{min}} = 0.01067733329246734

P_{\text{test}} = -5.1245 \times 10^{-5}

As P_{\text{min}} > P_{\text{test}} and Null hypothesis is accepted. Therefore in overall experiment there is no correlation between each set of real network gene ranks to random ranks.
Figure 38: Histogram of 1000 P-values resulting from correlation tests of real and permuted network ranks of genes relevant to Marfan syndrome group.

Correlation test for Pathways:

A similar histogram (Figure 39) is generated for all the pathways (from Biocarta, Kegg and Reactome pathway sources) by including their associated 3000 P-values correlations (1000 correlation P-values from each pathway source). Figure 39, illustrates that in only 5.1% times the null hypothesis is rejected. Meta-analysis is performed to test the significance of correlation for the entire pathways real ranks to random ranks.

\[ P_{min} = 7.43E-04 \]

\[ P_{test} = 1.538E-05 \]

as \( P_{min} > P_{test} \), Null hypothesis is accepted and concludes that in overall there is no correlation between each set of real network pathway ranks to random ranks.
Figure 39: Histogram of 1000 P-values resulting from correlation tests of real and permuted network ranks of pathways relevant to Marfan syndrome group

Correlation test for GO terms:

Correlation test was not performed on biological processes and molecular function type entities as random ranks hover over very few ranks and the data distribution is not compliant with this test.

Significance of entity based on score in real network:

This test is primarily to evaluate the significance of biological entity within the real network based on its score compared with random scores. From Table 8-3, It can be observed that all the associated entities are highly significant with relevant to this particular significance test. From their P-values it can be concluded that their scores in real network are higher compared to each score from 1000 randomized networks.
<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Id</th>
<th>Symbol</th>
<th>Real Score</th>
<th>Rank</th>
<th>P-value</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>2200</td>
<td>FBN1</td>
<td>205.2212</td>
<td>1</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Gene</td>
<td>7046</td>
<td>TGFBR1</td>
<td>193.8945</td>
<td>2</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Gene</td>
<td>7040</td>
<td>TGFBR1</td>
<td>160.5482</td>
<td>3</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Gene</td>
<td>7048</td>
<td>TGFBR2</td>
<td>141.686</td>
<td>4</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_tgfbPathway</td>
<td>TGF beta signaling pathway</td>
<td>1.465765</td>
<td>2</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Kegg)</td>
<td>hsa04350</td>
<td>TGF-beta signaling pathway</td>
<td>4.33663</td>
<td>4</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Reactome)</td>
<td>170834</td>
<td>TGF-beta signaling pathway</td>
<td>2.32222</td>
<td>1</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Biological Process (GO)</td>
<td>GO:0007179</td>
<td>Transforming growth factor beta receptor signaling pathway</td>
<td>3.411054</td>
<td>7</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Molecular Function (GO)</td>
<td>GO:0005201</td>
<td>Extracellular matrix structural constituent</td>
<td>10.52889</td>
<td>3</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
</tbody>
</table>

**Table 8-3:** Validation of importance of genomic entities based on their score in real network
<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Id</th>
<th>Symbol</th>
<th>Rank</th>
<th>P-value</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>2200</td>
<td>FBN1</td>
<td>1</td>
<td>0.367074</td>
<td>Significant</td>
</tr>
<tr>
<td>Gene</td>
<td>7046</td>
<td>TGFBR1</td>
<td>2</td>
<td>0.301337</td>
<td>Significant</td>
</tr>
<tr>
<td>Gene</td>
<td>7040</td>
<td>TGFB1</td>
<td>3</td>
<td>0.31011</td>
<td>Significant</td>
</tr>
<tr>
<td>Gene</td>
<td>7048</td>
<td>TGFBR2</td>
<td>4</td>
<td>0.191786</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h_tgfbPathway</td>
<td></td>
<td></td>
<td>2</td>
<td>0.451282</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Kegg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa04350</td>
<td></td>
<td></td>
<td>4</td>
<td>0.545022</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Reactome)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>170834</td>
<td></td>
<td></td>
<td>1</td>
<td>0.105099</td>
<td>Significant</td>
</tr>
<tr>
<td>Biological Process (GO)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0007179</td>
<td></td>
<td></td>
<td>7</td>
<td>1.06E-100</td>
<td>Significant</td>
</tr>
<tr>
<td>Molecular Function (GO)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005201</td>
<td></td>
<td></td>
<td>3</td>
<td>6.04E-204</td>
<td>Significant</td>
</tr>
</tbody>
</table>

**Table 8-4:** Probability and significance after applying Chi – Square test on each entity’s random ranks associated with Marfan

**Chi – Square method \( \chi^2 \)**

The higher P-values (p>0.05) associated with genes and pathways indicates that their random ranks are uniformly distributed (completely chaotic) and the real rank is significant. ‘**Bonferroni Correction**’ was not applied as the genomic entities are significant enough to conclude. Though, biological process ‘Transforming growth factor beta receptor signaling pathway’ (GO:0007179) and molecular function ‘Extracellular matrix structural constituent’ (GO:0005201) suffer from
extremely low P-values, the histogram (Figure 40) shows that the random ranks were not drifting around the real rank (GO:0007179- 7th rank, GO:0005201 – 3rd rank), thus still considering both these entities significant. The statistical analysis highlights that the ranked entities are in compliance with the real network and known to be involved with the phenome group under study.

**Z – Score:**
This test can’t be applied on genes as real scores and random scores are not compliant enough to perform the test. From Table 8-5, it can be depicted that all the entities associated with the Marfan syndrome group are highly significant.

**Case Study 2: Obesity Disease group**
Obesity is a rapidly becoming a serious concern of premature death and reaching epidemic proportions worldwide [146]. It increases the risk of cardiovascular diseases and insulin resistance, the core feature of type 2 diabetes even within the normal body mass index (BMI) range. Several cytokines and inflammatory markers contribute to the cardiovascular outcome in overweight and obese people[2]. American Heart Association (AHA) reclassified obesity as a ‘major, modifiable risk factor’ for coronary heart disease
Figure 40: Proof of these particular entities are significant though they have less Chi-square value (CHD)[147]. The insulin-resistant state of abdominal obesity adds substantially to the CHD risk of patients with familial hypercholesterolaemia[148]. Hypertension is approximately three times more common in obese than normal-weight persons[5]. Changes in the right heart also occur in obesity. The pathophysiology is related to obstructive sleep apnea and/or the obesity hypoventilation syndrome, which produce pulmonary hypertension and right ventricular hypertrophy, dilatation, progressive dysfunction, and finally failure[6, 149].
Table 8-5: significance of entities based on Z-score (Marfan syndrome group)

However, right ventricular dysfunction can also occur as a consequence of left ventricular dysfunction, and the heart failure that develops is often biventricular[7-9]. Figure 41 illustrates the obesity disease cluster and Table 8-6 gives the important biological entities associated with this group curated from literature. Computational predictions for this cluster are provided in Table 8-8 and it can be observed that 95% of literature curated facts are re-producible through our framework. Figure 42 provides the RDF visual depiction of ranked entities associated with Obesity group. The next few sections underlie the validity of our predictions through statistical analysis.
**Figure 41**: Obesity disease group curated from literature [1, 5-9]

<table>
<thead>
<tr>
<th>Functional Module Name</th>
<th>Supporting References (Pubmed Ids)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peroxisome Proliferators</td>
<td>9792666,10608261,16179270,17805473</td>
</tr>
<tr>
<td>Angiotensin II mediated activation</td>
<td>16806260,15276022,16389635</td>
</tr>
<tr>
<td>Role of PPAR-gamma Coactivators</td>
<td>16682454,12832613,17932317</td>
</tr>
<tr>
<td>Visceral Fat Deposits pathway</td>
<td>17884455,15181027,1794584</td>
</tr>
<tr>
<td>Angiotensin-converting enzyme 2 regulation</td>
<td>17537156,17520798,17401297,17899953</td>
</tr>
<tr>
<td>calcium/calmodulin-dependent kinase</td>
<td>3138915,8635670,10630371,2387254</td>
</tr>
<tr>
<td>Adhesion Molecules lymphocyte</td>
<td>14968297,10675271,17718717</td>
</tr>
<tr>
<td>Wnt Signalling</td>
<td>17032748,16723389,16723389</td>
</tr>
<tr>
<td>LR P6 Signalling</td>
<td>17633795,17332414</td>
</tr>
<tr>
<td>Nitric Oxide in the Heart</td>
<td>17941867,17933859</td>
</tr>
<tr>
<td>Angiogenesis</td>
<td>17804301,17906444,17786229</td>
</tr>
<tr>
<td>Functional Role in Vascular</td>
<td></td>
</tr>
<tr>
<td>Endothelium</td>
<td>17940464,17933859,17697061</td>
</tr>
<tr>
<td>Neutrophil</td>
<td>17697874,16537681,17697874</td>
</tr>
<tr>
<td>Bone Mineralization</td>
<td>17939041,17095733,16601233</td>
</tr>
<tr>
<td>Low-density lipoprotein (LDL) pathway</td>
<td>17583182,17437081,16272194</td>
</tr>
<tr>
<td>p38 MAPK Signaling Pathway</td>
<td>17940160,17702846,17554073</td>
</tr>
<tr>
<td>Calcium signaling pathway</td>
<td>17130640,15831571,16966319</td>
</tr>
<tr>
<td>Focal adhesion</td>
<td>17132824,10905491,17394460</td>
</tr>
<tr>
<td>Adipocytokine signaling</td>
<td>17467106,15558058,17615379,16847434</td>
</tr>
<tr>
<td>TGF-beta signaling</td>
<td>17786229,17686962,15928193,9636194</td>
</tr>
<tr>
<td>Hematopoietic cell lineage</td>
<td>12849006,8616721,14607907</td>
</tr>
<tr>
<td>Cell adhesion molecules (CAMs)</td>
<td>17613807,17511629,17920663</td>
</tr>
<tr>
<td>Linoleic acid</td>
<td>17878672,17168664,17940463</td>
</tr>
<tr>
<td>Purine metabolism</td>
<td>16375734,15671926,2805359</td>
</tr>
<tr>
<td>Riboflavin metabolism,</td>
<td>16806731,17412321</td>
</tr>
<tr>
<td>CoA biosynthesis</td>
<td>17468937,15533614,17126822,16685502</td>
</tr>
<tr>
<td>Pantothenate</td>
<td>17639559</td>
</tr>
<tr>
<td>nicotinamide metabolism</td>
<td>1584966,15456332,15670660,16434553</td>
</tr>
<tr>
<td>Glycerolipid metabolism</td>
<td>15102885,16843721,17389595</td>
</tr>
<tr>
<td>Insulin receptor mediated signaling</td>
<td>17914241,17805473</td>
</tr>
<tr>
<td>Xenobiotic metabolism</td>
<td>7713315,16424281</td>
</tr>
</tbody>
</table>
Table 8-6: Literature curated important biological entities implicated in Obesity group

Statistical validation

Table 8-7 provides the total number of entities for each genomic type associated with the obesity group. We randomized the genome network associated with obesity group for 1000 times and performed statistical tests to verify the significance of the biological entities curated from literature (genes, pathways, biological processes and molecular functions). We couldn’t find any genes from our data set in literatures that are pathophysiologically associated with the obesity disease group and statistical tests were ignored on gene set.
**Figure 42**: RDF visualizer of ranked entities associated with Obesity group

**Correlation test**

Sample correlations are calculated between each set of real ranks and random ranks per each randomization both for genes and pathways (1000 randomizations).

**Correlation test for Pathways**

A similar histogram (Figure 43) is generated for all the pathways (from Biocarta, Kegg and Reactome pathway sources) by including their associated 3000 P-values correlations (1000 correlation P-values from each pathway source).
<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Total Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>246</td>
</tr>
<tr>
<td>Pathway (BIOCARTA)</td>
<td>142</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>55</td>
</tr>
<tr>
<td>Pathway (REACTOME)</td>
<td>13</td>
</tr>
<tr>
<td>Biological Process (GO)</td>
<td>197</td>
</tr>
<tr>
<td>Molecular Function (GO)</td>
<td>160</td>
</tr>
</tbody>
</table>

Table 8-7: Statistics of different total Entities associated with Obesity group

Figure 43 illustrates that in only 1.26% of times the null hypothesis is rejected. Therefore, meta-analysis is performed to test the significance of correlation for the entire pathways real ranks to random ranks.

\[ P_{\text{min}} = 0.005697487 \]

\[ P_{\text{test}} = 1.538 \times 10^{-5} \]

as \( P_{\text{min}} > P_{\text{test}} \), Null hypothesis is accepted and concludes that in overall there is no correlation between each set of real network pathway ranks to random ranks.
Figure 43: Histogram of 1000 P-values resulting from correlation tests of real and permuted network ranks of pathways relevant to Obesity phenotype group.

Correlation test for GO terms

Correlation was not performed on biological processes and molecular function type genomic entities as random ranks hover over very few ranks.

Significance of entity based on score in real network

This test is primarily to evaluate the significance of biological entity within the real network based on its score compared with random scores. From Table 8-8, it can be observed that all the associated entities except Reactome pathway ‘Insulin receptor mediated signaling (REACTOME_74752)’ and molecular function ‘melanocortin receptor activity (GO_0004977)’ is highly significant with relevant to this particular significance test. The P-values conclude that their scores in real network are higher (from P-value) compared to each score from 1000 randomized networks.
Chi – Square method \( \chi^2 \)

The higher P-values (p>0.05) associated with pathways (From Table 8-8) indicates that their random ranks are uniformly distributed (completely chaotic) and the real rank is significant. As entities are already having required P-value to prove their significance, we did not further calculate ‘Bonferroni Correction’. The following biological processes

- GO:0006813 - potassium ion transport
- GO:0006954 - inflammatory response
- GO:0007186 - G-protein coupled receptor protein signaling pathway
- GO:0006629 - lipid metabolism

are extremely insignificant from their P-values, but from the Figure 44, it can be seen that their real ranks are not in proximity of the random ranks and still can be considered significant. All the ‘molecular functions’ suffer from extremely low P-values, but the histogram (Figure 45) reveals that the real ranks were not drifting in the vicinity of the random ranks thus still considering these entities significant.
Table 8-8: Significance of various genomic entities associated with Obesity disease group from various statistical tests

<table>
<thead>
<tr>
<th>Source</th>
<th>Id</th>
<th>Description</th>
<th>RealRank</th>
<th>Test on Real network Scores (P-Value)</th>
<th>Significance Level</th>
<th>Chi-Square Test (P-value)</th>
<th>Significance Level (P-value)</th>
<th>Z-Score (P-value)</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathway</td>
<td>h_pparaPathway</td>
<td>Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha)</td>
<td>1</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>0.1487889</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway</td>
<td>h_At1rPathway</td>
<td>Angiotensin II mediated activation of JNK Pathway via Pyk2 dependent signaling</td>
<td>2</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>0.2740438</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway</td>
<td>h_ppargPathway</td>
<td>Role of PPAR-gamma Coactivators in Obesity and Thermogenesis</td>
<td>3</td>
<td>&lt;0.0011</td>
<td>Significant</td>
<td>0.244935</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway</td>
<td>h_vobesityPathway</td>
<td>Visceral Fat Deposits and the Metabolic Syndrome</td>
<td>5</td>
<td>&lt;0.0011</td>
<td>Significant</td>
<td>0.2414362</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway</td>
<td>h_ace2Pathway</td>
<td>Angiotensin-converting enzyme 2 regulates heart function</td>
<td>6</td>
<td>&lt;0.0011</td>
<td>Significant</td>
<td>0.9939544</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway</td>
<td>h_hifPathway</td>
<td>Hypoxia-Inducible Factor in the Cardiovascular System</td>
<td>8</td>
<td>&lt;0.0011</td>
<td>Significant</td>
<td>0.7965146</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway</td>
<td>h_hdacPathway</td>
<td>Control of skeletal myogenesis by HDAC &amp; calcium/calmodulin-dependent kinase (CaMK)</td>
<td>12</td>
<td>0.008</td>
<td>Significant</td>
<td>0.4313268</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway</td>
<td>h_lymphocytePathway</td>
<td>Adhesion Molecules on Lymphocyte</td>
<td>13</td>
<td>0.009</td>
<td>Significant</td>
<td>0.1988451</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Source</td>
<td>Id</td>
<td>Description</td>
<td>RealRank</td>
<td>Test on Real network Scores (P-Value)</td>
<td>Significance Level</td>
<td>Chi-Square Test (P-value)</td>
<td>Significance Level (P-value)</td>
<td>Z-Score (P-value)</td>
<td>Significance Level</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------------</td>
<td>--------------------------------------------------</td>
<td>----------</td>
<td>--------------------------------------</td>
<td>-------------------</td>
<td>--------------------------</td>
<td>-----------------------------</td>
<td>-----------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_wnt-lrp6Pathway</td>
<td>Wnt/LRP6 Signalling</td>
<td>16</td>
<td>0.012</td>
<td>Significant</td>
<td>0.4642387</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_no1Pathway</td>
<td>Actions of Nitric Oxide in the Heart</td>
<td>17</td>
<td>0.009</td>
<td>Significant</td>
<td>0.3295415</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_vegfPathway</td>
<td>VEGF, Hypoxia, and Angiogenesis</td>
<td>17</td>
<td>0.015</td>
<td>Significant</td>
<td>0.3212777</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_raccPathway</td>
<td>Ion Channels and Their Functional Role in Vascular Endothelium</td>
<td>17</td>
<td>0.008</td>
<td>Significant</td>
<td>0.1513021</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_neutrophilPathway</td>
<td>Neutrophil and Its Surface Molecules</td>
<td>18</td>
<td>0.009</td>
<td>Significant</td>
<td>0.1253309</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_npp1Pathway</td>
<td>Regulators of Bone Mineralization</td>
<td>19</td>
<td>0.009</td>
<td>Significant</td>
<td>0.3379105</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_LDLpathway</td>
<td>Low-density lipoprotein (LDL) pathway during atherogenesis</td>
<td>21</td>
<td>0.01</td>
<td>Significant</td>
<td>0.1897858</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (Biocarta)</td>
<td>h_p38mapkPathway</td>
<td>p38 MAPK Signaling Pathway</td>
<td>22</td>
<td>0.014</td>
<td>Significant</td>
<td>0.807635</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa04020</td>
<td>Calcium signaling pathway</td>
<td>2</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>0.4394369</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa04510</td>
<td>Focal adhesion</td>
<td>3</td>
<td>0.001</td>
<td>Significant</td>
<td>0.4443191</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa04920</td>
<td>Adipocytokine signaling pathway</td>
<td>7</td>
<td>0.006</td>
<td>Significant</td>
<td>0.4454363</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa04310</td>
<td>Wnt signaling pathway</td>
<td>8</td>
<td>0.015</td>
<td>Significant</td>
<td>0.8709276</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa04350</td>
<td>TGF-beta signaling pathway</td>
<td>9</td>
<td>0.03</td>
<td>Significant</td>
<td>0.3652021</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa04640</td>
<td>Hematopoietic cell lineage</td>
<td>10</td>
<td>0.031</td>
<td>Significant</td>
<td>0.116032</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Source</td>
<td>Id</td>
<td>Description</td>
<td>RealRank</td>
<td>Test on Real network Scores (P-Value)</td>
<td>Significance Level</td>
<td>Chi-Square Test (P-value)</td>
<td>Significance Level (P-value)</td>
<td>Z-Score (P-value)</td>
<td>Significance Level</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------</td>
<td>-------------------------------------------</td>
<td>----------</td>
<td>--------------------------------------</td>
<td>--------------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa04514</td>
<td>Cell adhesion molecules (CAMs)</td>
<td>11</td>
<td>0.025</td>
<td>Significant</td>
<td>0.2326947</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa00590</td>
<td>Arachidonic acid metabolism</td>
<td>13</td>
<td>0.029</td>
<td>Significant</td>
<td>0.7031609</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa00591</td>
<td>Linoleic acid metabolism</td>
<td>14</td>
<td>0.034</td>
<td>Significant</td>
<td>0.9661938</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa00230</td>
<td>Purine metabolism</td>
<td>15</td>
<td>0.026</td>
<td>Significant</td>
<td>0.8086631</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa00740</td>
<td>Riboflavin metabolism</td>
<td>15</td>
<td>0.034</td>
<td>Significant</td>
<td>0.6517661</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa00770</td>
<td>Pantothenate and CoA biosynthesis</td>
<td>15</td>
<td>0.029</td>
<td>Significant</td>
<td>0.534651</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa00760</td>
<td>Nicotinate and nicotinamide metabolism</td>
<td>15</td>
<td>0.024</td>
<td>Significant</td>
<td>0.97298</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (KEGG)</td>
<td>hsa00561</td>
<td>Glycerolipid metabolism</td>
<td>16</td>
<td>0.026</td>
<td>Significant</td>
<td>0.4131139</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (REACTOME)</td>
<td>REACTOME_73923</td>
<td>Lipid metabolism</td>
<td>2</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>0.0548234</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>Pathway (REACTOME)</td>
<td>REACTOME_74752</td>
<td>Insulin receptor mediated signaling</td>
<td>5</td>
<td>0.642</td>
<td>Insignificant</td>
<td>0.2562154</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0006813</td>
<td>potassium ion transport</td>
<td>5</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0006954</td>
<td>inflammatory response</td>
<td>7</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0007186</td>
<td>G-protein coupled receptor protein signaling pathway</td>
<td>8</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0006629</td>
<td>lipid metabolism</td>
<td>9</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0007190</td>
<td>adenylate cyclase activation</td>
<td>12</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Insignificant</td>
<td>&lt;0.0010</td>
<td>Insignificant</td>
</tr>
<tr>
<td>Source</td>
<td>Id</td>
<td>Description</td>
<td>RealRank</td>
<td>Test on Real network Scores (P-Value)</td>
<td>Significance Level</td>
<td>Chi-Square Test (P-value)</td>
<td>Significance Level (P-value)</td>
<td>Z-Score (P-value)</td>
<td>Significance Level</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>----------</td>
<td>--------------------------------------</td>
<td>--------------------</td>
<td>--------------------------</td>
<td>-----------------------------</td>
<td>---------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0007188</td>
<td>G-protein signaling, coupled to cAMP nucleotide second messenger</td>
<td>17</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Insignificant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0008203</td>
<td>cholesterol metabolism</td>
<td>21</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Insignificant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0007179</td>
<td>transforming growth factor beta receptor signaling pathway</td>
<td>24</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Insignificant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - BP</td>
<td>GO_0045843</td>
<td>negative regulation of striated muscle development</td>
<td>25</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Insignificant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0005524</td>
<td>ATP binding</td>
<td>1</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0001584</td>
<td>rhodopsin-like receptor activity</td>
<td>8</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0005509</td>
<td>calcium ion binding</td>
<td>10</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0030955</td>
<td>potassium ion binding</td>
<td>12</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0005125</td>
<td>cytokine activity</td>
<td>15</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0004674</td>
<td>protein serine/threonine kinase activity</td>
<td>16</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0020037</td>
<td>heme binding</td>
<td>18</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0003707</td>
<td>steroid hormone receptor activity</td>
<td>20</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0005024</td>
<td>transforming growth factor beta receptor activity</td>
<td>21</td>
<td>0.049</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0005516</td>
<td>calmodulin binding</td>
<td>22</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
<tr>
<td>GO - MF</td>
<td>GO_0004977</td>
<td>melanocortin receptor activity</td>
<td>25</td>
<td>0.872</td>
<td>Insignificant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
<td>&lt;0.0010</td>
<td>Significant</td>
</tr>
</tbody>
</table>
Figure 44: Proof of these particular GO-Biological processes associated with obesity group are significant though they have less Chi-square value.

Figure 45: Proof of these particular GO-Mol Functions associated with obesity group are significant though they have less Chi-square value.

Z-Score

All the P-values from Z-Score test statistics (Table 8-8) signify that the scores of all entities from real network are greater than the average of random scores. This provides additional
evidence of the importance of these entities in relevance to the relations of other entities in real network.

From the majority of conducted statistical tests highlights that the real entities ranks are in compliance with the real network and computationally validate our literature findings cohort with the phenome group under study. Though ranked clinical feature entities are presented in the RDF graphs (Figure 37, Figure 42), we did not provide any statistical tests for these as in our case we are looking to find and validate functional modules. The clinical features merely contribute significance to the genes they are associated with. But as the whole information space is ranked even the RDF can be probed for ranked clinical features, another significant advantage of using Semantic Web. Figure 46 explains the significance of using feed –back[40] based scoring of each graph entity and also realizes our earlier assumption that pathway ranks are based on importance of genes participating in that pathway. ‘Angiotensin-converting enzyme 2 regulates heart function’ (h_ace2Pathway), a Biocarta Pathway, is ranked high in the Obesity associated RDF graph though it has only two genes (AGT and AGTR1) from our gene data set. A careful analysis of the graph (Figure 46) for high ranking of this pathway explains the high ranked genes participating in that pathway (Figure 8-16). It also explains that though ‘Bioactive Peptide Induced Signaling Pathway’ (from Biocarta) has more genes, but still ranked (rank: 9) low compared to ‘h_ace2Pathway’ (Rank: 6) due to less important genes.
**Figure 46**: Obesity associated graph depicting that ‘*Angiotensin-converting enzyme 2 regulates heart function*’ is ranked high due to its high scored two participating genes (AGT, AGTR1) from our data set.
Summary and Future Directions

Summary

Diseases may not be as independent of each other as medical practitioners currently consider them to be. Though the treatment often varies, most human diseases are dependent sharing multiple clinical symptoms of each other. Many phenotypes at microscopic level are associated with multiple levels of functional modules that can be best described as sub-networks of a complex network. These sub-networks connect many cellular components and an understanding of the functionally relevant genetic, regulatory, and protein – protein interactions in a cellular network will play an important role in discerning the pathophysiology of human diseases [150]. Disease mechanisms can be understood by trying to map the implicated genes and gene products into detailed wiring diagram of various influenced cellular components. Such kinds of network based approaches are already providing optimal solutions in discerning the pathogenic biological entities[21, 93, 137, 150]. Our approach provides a flexible and powerful framework to achieve solutions for these sorts of specific problems. Here we emphasize on combination and extension of different widely used methods, which with certain order can leverage on their strengths to approach the problem solvable domain.

Sub-cellular components and disease genes are linked by complex molecular links. Any computational algorithm being developed has to consider these intricate relationships in order to dissect disease mechanism. RDF, a Semantic Web standard perfectly suits to establish and replicate the intra cellular networks by providing the required semantics and formal knowledge representation. As RDF is already a graph, any required graph
algorithms can be extended to mine this graph. Here, we have shown how Semantic Web standards can be used to aggregates broad sets of data, and prioritizing them for relative contribution and relevance using a page-ranking approach. In general, we are applying network centrality analysis to rank resources according to their importance within the Bio-RDF network structure. Moreover, here the importance of a resource is calculated by diverse data sets (from genome to phenome) integrated into the information space. Additionally, resource ranking is performed semantically by including contextual semantic weights on the properties connecting the resources. Results from case studies indicate the value of data mining of our cross connective phenome similarity and functional gene network across multiple heterogeneous data sets in prioritizing disease relevant entities.

Indeed, the recent published work by the The Wellcome Trust Case Control Consortium[151] illustrates that genome wide approaches require large sampling of molecular clinical data (polymorphisms, symptoms), and yield many more gene associations than previously obtained from earlier limited diseases studies. We have shown how semantically rich information of symptoms can be used to relate their causative diseases to underlying genetic components. This differs substantially by work of others, such as Goh et al [93], who used OMIM information directly to link genes with diseases. Our approach has the advantage that we utilize the breadth of symptomatic evidence to establish possible associations between different diseases, there by not relying solely on our current limited knowledge of the genetic basis. We see our approach being able to accommodate the inherent noise in the information sets, whether it comes from improper classifications, speculative hypotheses, or faulty reasoning. As more evidence and interpretations get compiled, we anticipate broad semantic analysis becoming more robust
and yielding more insights. We believe the use of diverse sets of evidence and hypotheses will greatly advance the set of testable models for overlapping diseases mechanisms. This will consequently have a direct impact in the development of both new and re-directed therapeutic applications. The key requirement we propose is that the application of semantics in data sets and curated interpretations help to manage and analyze the vast biomedical resources being generated, and will become an indispensable tool for biomedical researchers.

Apart from coverage of data sets used, we have shown how we can leverage on Semantic Web standards and techniques to apply on specific biological problem, right from RDF and OWL for integration, application of customized network centrality analysis algorithms for mining Bio-RDF and also retrieving ranked results using graph query languages such as SPARQL. We did not find any other similar frame work using Semantic web technologies to create and integrate phenomics and genomics knowledge into a network platform to mine and prioritize biological entities. Although, in the current study we focused on the cardiovascular system, our approach can be applied to any group of genes or disease sets. One immediate application could be in applying to OMIM diseases (around 1554) having known loci but unknown molecular basis. This in turn will enable targeted research on how mutations in the gene contribute to disease and provide specific leads towards novel diagnostic and therapeutic approaches. Our framework builds on the concept of disease modularity [14, 19, 152], which is invaluable for pharmaceutical industry for identifying therapeutic targets or combinations of targets that can alter disease expression and especially in drug repositioning. Finally, we strongly believe that our success of methods can be attributed to a combination of factors and diverse implemented
techniques from the integration of functional genomic data with a phenotype similarity scheme, thereby taking advantage of the complete clinical feature spectrum of cardiovascular syndrome related human diseases did accelerate the disease pathogenesis discovery process by gathering and sifting through all available –omics knowledge.

**Future Directions**

There are multiple directions in which this project can be improved and extended. One immediate application could be to mine all the OMIM phenotypes in order to include all the diseases as here we restricted to cardiovascular disease domain. Apart from using PCA for dimensionality reduction, the better and more appropriate approach would be using graph clustering methods. Instead of converting a phenome similarity matrix into a RDF network, an ideal approach would be to extend and apply graph clustering algorithms onto the RDF. Due to inherent semantics buried in RDF, it provides more flexibility in controlling the semantic parameters for clustering. A graph clustering approach might not only group diseases but also expose the relevant clinical features associated with each cluster which might not be possible using traditional clustering algorithms as they consider the clinical feature space for clustering. Finally, pheno matrix is always sparse and graph algorithms seem to be more appropriate for such data features. Apart from graph clustering an ontology based approach provides another alternative for improved disease nosology. One interesting area would be using OWL – DL (OWL- Description logics) standard to classify diseases based on pure knowledge based reasoning from a fixed vocabulary such as UMLS which annotates diseases.
Right now the bridge connecting phenome and genome are only genes which are quite few with limited data sources providing such relations. Mining disease – gene associations from literature will help to provide much more in-depth layer for computational accuracy and new discovery. Not only limiting to genes as the bridging points but also expanding those to pathways, protein complexes, process and functions will provide much more inherent robust framework to extend functional predictions to other genomic elements. Extending the network to further include pharmacome data will enhance and extend our current application framework to phenome – genome - pharmacome network. This will be greatly beneficial to the pharmaceutical industry to enhance new knowledge discovery for understanding drug targets in the context of genome and phenome networks.
Bibliography


[60] "GOA," [http://www.ebi.ac.uk/GOA/](http://www.ebi.ac.uk/GOA/).


"r-Project," [http://www.r-project.org](http://www.r-project.org).


A. Moustafa, "JAligner: Open source Java implementation of Smith-Waterman."


Appendix A – Oracle 10g RDF Store Data Structure

Color index

Table : GENE_INFO
Model : GENE_INFO
Converted raw data into RDF triples using the below RDF model
Table: GENEONTOLOGY
Model: GENEONTOLOGY

Table: GOA
Model: GOA
Converted raw data into RDF triples using the below RDF model
**Table**: INTERACTIONS  
**Model**: INTERACTIONS  

---

**Table**: PATHWAYS  
**Model**: PATHWAYS  
**Data**: Downloaded from the following data sources.  
- Biocarta Pathways: www.biocarta.com/  
- Kegg Pathways: www.genome.ad.jp/kegg/pathway.html  
- Reactome Pathways: http://www.reactome.org/  
Converted raw data into RDF triples using the below RDF model
**Table**: OMIM  
**Model**: OMIM  
**Data**: Downloaded from the following data sources.  
- Disease - Concepts: umlsks.nlm.nih.gov/  

Converted raw data into RDF triples using the below RDF model

**Table**: MOUSE_PHENOTYPE  
**Model**: MOUSE_PHENOTYPE  
**Data**: Downloaded from Jackson labs FTP [www.jax.org/](http://www.jax.org/)  
Converted raw data into RDF triples using the below RDF model
Appendix B - Sample RDF/XML of Phenome and Genome RDF network

Sample Genome RDF network in RDF/XML for BRCA2

<DCO:Gene rdf:ID="GENE_675">
  <DCO:hasDescription>Breast cancer 2, early onset</DCO:hasDescription>
  <rdfs:label>BRCA2</rdfs:label>
  <DCO:hasMapLocation>13q12.3</DCO:hasMapLocation>
  <DCO:onChromosome>13</DCO:onChromosome>
  <DCO:inMolecularFunction rdf:resource="#GO_0003676">
    <DCO:GO_MolecularFunction rdf:ID="GO_0003676">
      <DCO:dbXrefTo>
        <rdfs:label>nucleic acid binding</rdfs:label>
      </DCO:dbXrefTo>
    </DCO:GO_MolecularFunction>
  </DCO:inMolecularFunction>
  <DCO:inBiologicalProcess rdf:resource="#GO_0007093">
    <DCO:GO_BiologicalProcess rdf:ID="GO_0007093">
      <DCO:dbXrefTo>
        <rdfs:label>mitotic checkpoint</rdfs:label>
      </DCO:dbXrefTo>
    </DCO:GO_BiologicalProcess>
  </DCO:inBiologicalProcess>
  <DCO:inCellularComponent rdf:resource="#GO_0030141">
    <DCO:GO_CellularComponent rdf:ID="GO_0030141">
      <DCO:dbXrefTo>
        <rdfs:label>secretory granule</rdfs:label>
      </DCO:dbXrefTo>
    </DCO:GO_CellularComponent>
  </DCO:inCellularComponent>
  <DCO:participates_inReactomePathway rdf:resource="#REACTOME_73894">
    <DCO:Pathway rdf:ID="REACTOME_73894">
      <rdfs:label>DNA Repair</rdfs:label>
    </DCO:Pathway>
  </DCO:participates_inReactomePathway>
  <DCO:participates_inBiocartaPathway rdf:resource="#h_atrbrcaPathway">
    <DCO:Pathway rdf:ID="h_atrbrcaPathway">
      <rdfs:label>Role of BRCA1, BRCA2 and ATR in Cancer Susceptibility</rdfs:label>
    </DCO:Pathway>
  </DCO:participates_inBiocartaPathway>
  <DCO:interacts_with rdf:resource="#GENE_2316">
    <DCO:Gene rdf:ID="GENE_2316">
      <DCO:hasMapLocation>Xq28</DCO:hasMapLocation>
      <DCO:onChromosome>X</DCO:onChromosome>
      <DCO:hasDescription>filamin A, alpha</DCO:hasDescription>
      <rdfs:label>FLNA</rdfs:label>
    </DCO:Gene>
  </DCO:interacts_with>
</DCO:Gene>
Sample Phenome RDF network in RDF/XML for SHPRINTZEN-GOLDBERG CRANIOSYNOSTOSIS SYNDROME

```xml
<DCO:Disease rdf:ID="OMIM_182212">
  <rdfs:label>SHPRINTZEN-GOLDBERG CRANIOSYNOSTOSIS SYNDROME NIDDM1</rdfs:label>
  <DCO:has_alternative_label>CRANIOSYNOSTOSIS WITH ARACHNODACTYLY AND MARFANOID DISORDER WITH CRANIOSYNOSTOSIS, TYPE I</DCO:has_alternative_label>
  <DCO:has_alternative_label>MARFANOID CRANIOSYNOSTOSIS SYNDROME</DCO:has_alternative_label>
  <DCO:hasBioResource>
    <DCO:BioResource rdf:ID="OMIM"/>
  </DCO:hasBioResource>
  <DCO:hasAssociatedGene>
    <DCO:Gene rdf:ID="GENE_2200"/>
  </DCO:hasAssociatedGene>
  <DCO:hasAffectedAnatomy>
    <DCO:Clinical_Attribute rdf:ID="CUI_C0013273">
      <rdfs:label>Structure of ductus arteriosus</rdfs:label>
    </DCO:Clinical_Attribute>
  </DCO:hasAffectedAnatomy>
  <DCO:hasClinicalSymptom>
    <DCO:Clinical_Attribute rdf:ID="CUI_C0333706">
      <rdfs:label>Pericentric inversion</rdfs:label>
    </DCO:Clinical_Attribute>
  </DCO:hasClinicalSymptom>
  <DCO:hasPhenotypeSimilarityRelation>
    <DCO:PhenotypeSimilarityRelation rdf:about="#182212_154700">
      <DCO:clinically_similarTo>
        <DCO:Disease rdf:ID="OMIM_154700"/>
      </DCO:clinically_similarTo>
      <DCO:similarityValue rdf:datatype="http://www.w3.org/2001/XMLSchema#double">0.86464283045498</DCO:similarityValue>
    </DCO:PhenotypeSimilarityRelation>
  </DCO:hasPhenotypeSimilarityRelation>
</DCO:Disease>
```