A Genome – Phenome Integrated Approach for Mining Disease-Causal Genes using Semantic Web

Ranga Chandra Gudivada[1,3], Xiaoyan A. Qu [1,3], Anil G Jegga[2,3], Eric K. Neumann[4] , Bruce J Aronow[1,2,3]

Departments of Biomedical Engineering[1] and Pediatrics[2], University of Cincinnati and Division of Biomedical Informatics[3], Cincinnati Childrens Hospital Medical Center, Cincinnati OH-45229, USA, Teranode Corporation [4], Seattle, WA 98104

ABSTRACT

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.

Categories and Subject Descriptors

J.3 [Computer Applications] LIFE AND MEDICAL SCIENCES - *Biology and genetics*; E.1 [Data] DATA STRUCTURES - *Graphs and networks*; H.3.3 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval; I.2.4 [ARTIFICIAL INTELLIGENCE] Knowledge Representation Formalisms and Methods.

General Terms

Algorithms, Standardization, Languages

Keywords

Semantic Web, RDF, OWL, SPARQL, Semantic Ranking, Ontologies, Data Integration, Bioinformatics, UMLS

Correspondence

Gudivada Ranga Chandra, Email : gudx6u@cchmc.org

Bruce Aronow, Email : Bruce.Aronow@cchmc.org

WWW 2007, May 8--12, 2007, Banff, Canada.

1. INTRODUCTION

The identification of genes responsible for human disease is critical to comprehend 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 to several hundred genes. Similarly, DNA microarray experiments generate hundreds of differentially expressed genes. Apparently, both of these strategies fail to help researchers in reducing the target genes to a manageable number or to prioritize the disease specific causal genes for further analysis. This explains the need to develop sophisticated techniques and tools to identify key candidates from gene lists generated by disease gene discovery methods.

Disease gene discovery has been shown to be accelerated by applying aggregative computational methodologies on integrated data sets generated from genome-scale experiments [1]. Integrating diverse functional genomic data has several advantages as described by Giallourakis et al [2]. 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 [1, 3-10]. Some of them [1, 3, 4] 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 [5, 7, 10] 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) [11] 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 [12] of *Kleinberg algorithm* [13] 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 hypothesis [14]. Apart from *Kleinberg Authoritative Scores* , there are several other centrality measurements, such as *Google PageRank* [15], *Centrality Indices* [16] and *C-F closeness* [17]. At present, SW querying languages do not rank the retrieved results from RDF graphs, so we borrowed a technique from M. Sougata et al [12, 18] 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[1, 10]. 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 [19] and (b) genes differentially expressed in dilated cardiomyopathy [20].

2. METHODS

2.1 Data Sources

We used both genomic and phenomic data sources to prioritize gene candidates (See Figure 1).

Genomic Data Sources

1) Gene Ontology (GO) [21] was downloaded from GeneOntology website (geneontology.org/ontology/ gene_ontology_edit.obo). Corresponding human GO-gene annotations were downloaded from NCBI Entrez Gene ftp site (ftp.ncbi.nih.gov/gene/DATA/gene2go.gz). The resultant data set contained 15068 human genes annotated with 7124 unique GO terms.
2) Gene-pathway annotations were compiled from KEGG [22], BioCarta (http://www.biocarta.com/), BioCyc [23] and Reactome [24] . 4772 human genes had at least one pathway association (a total of 672 pathways).

Phenomic Data Sources

1) Mammalian Phenotype (MP) ontology [25] 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.
2) Online Mendelian Inheritance in Man (OMIM) [26] was searched for the '*cardiovascular*' phrase occurring in Clinical Synopsis (CS) section and a total of 936 records were downloaded in XML format. OMIM ID and the corresponding gene associations were downloaded from NCBI Entrez Gene ftp site (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/ mim2gene).
3) 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 [27] 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.

2.2 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) concepts using MetaMap [26]. 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. 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 (Table 1). Clinical Features not having a corresponding CUI were given a custom unique id.

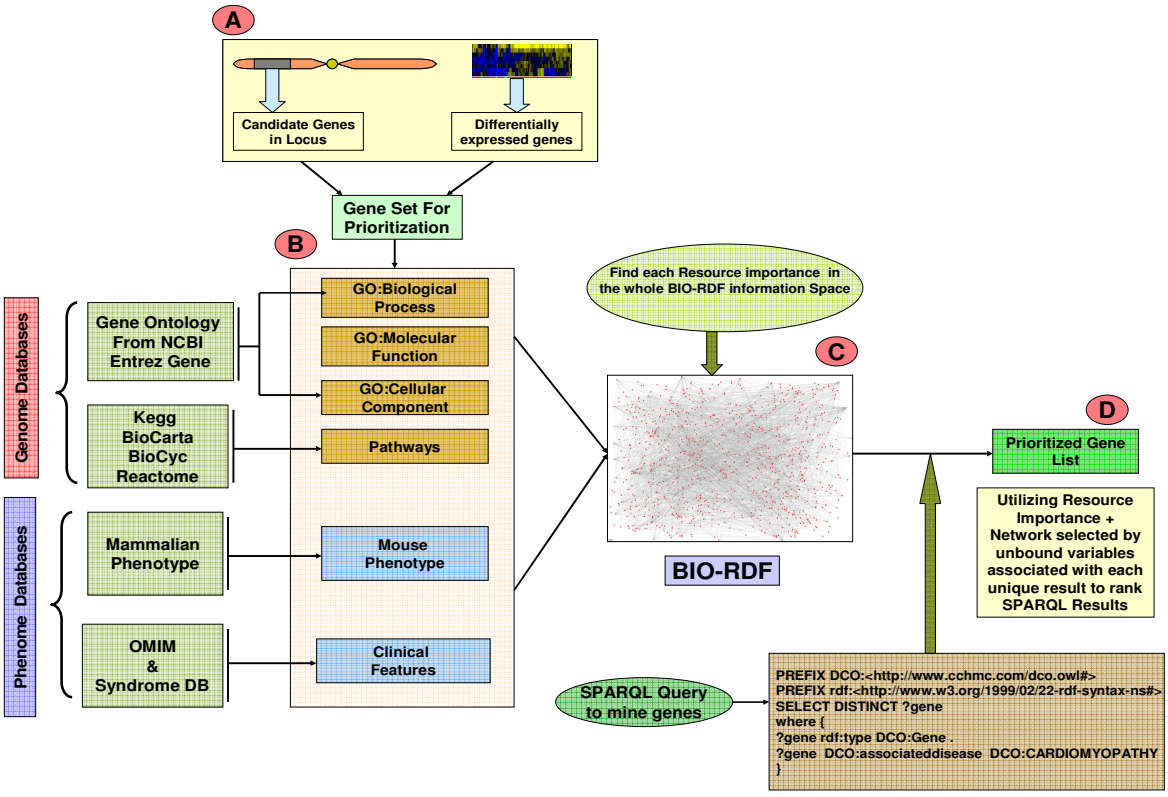

Figure 1: Schema diagram. A) Test gene set is obtained from a locus identified by linkage analysis, or a differentially expressed gene set from a microarray experiment. B) Genome and Phenome databases considered to create Bio-RDF includes GO: Molecular Function, GO: Biological Process, pathway, Mammalian 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.

Table 1. Statistics of clinical features from CS (OMIM) and MF (Syndrome db) mapped to UMLS concepts.

	Clinical Features parsed	Clinical Features Having CUI	Total Mapped CUI
OMIM	6838	4336	2660
Syndrome db	3887	2047	1332

2.3 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*' dataset.

2.4 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 [28]. 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 [26] in OWL, a language layered on top of RDF to offer support for axioms and inference. Jena (www.jena.sourceforge.net), a java frame work 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

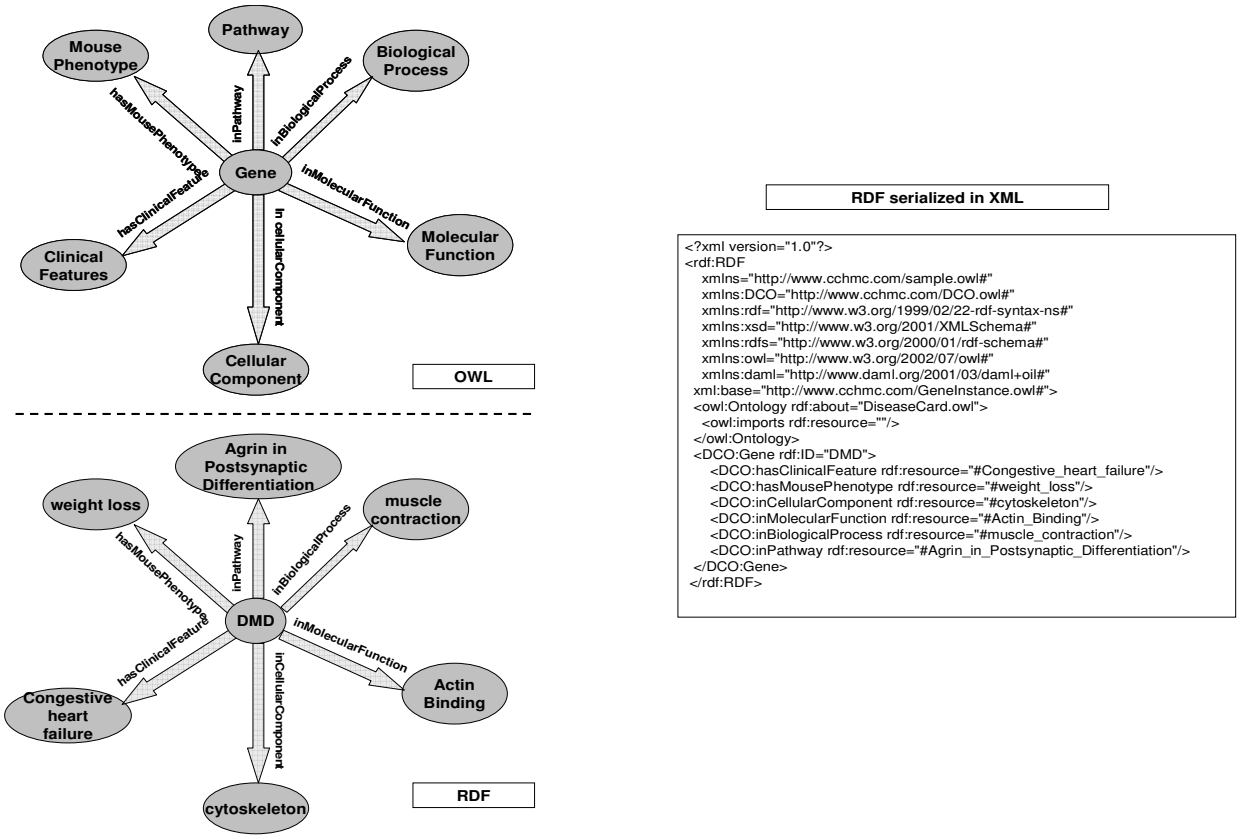

Figure 2: Portion of Bio-RDF for gene DMD based on the DCO ontology. The upper network is the ontology providing the required semantics for the lower RDF network consisting instance data

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-1). Figure-2 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.

2.5 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 [18]. 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 [12, 18].

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 [13] *hub* and *authority scores* for the WWW graph (Figure-3). 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. Figure-4 illustrates the significance of semantic weights on gene - pathway association. On the other hand, property *associated_process* links mouse phenotype to biological process.

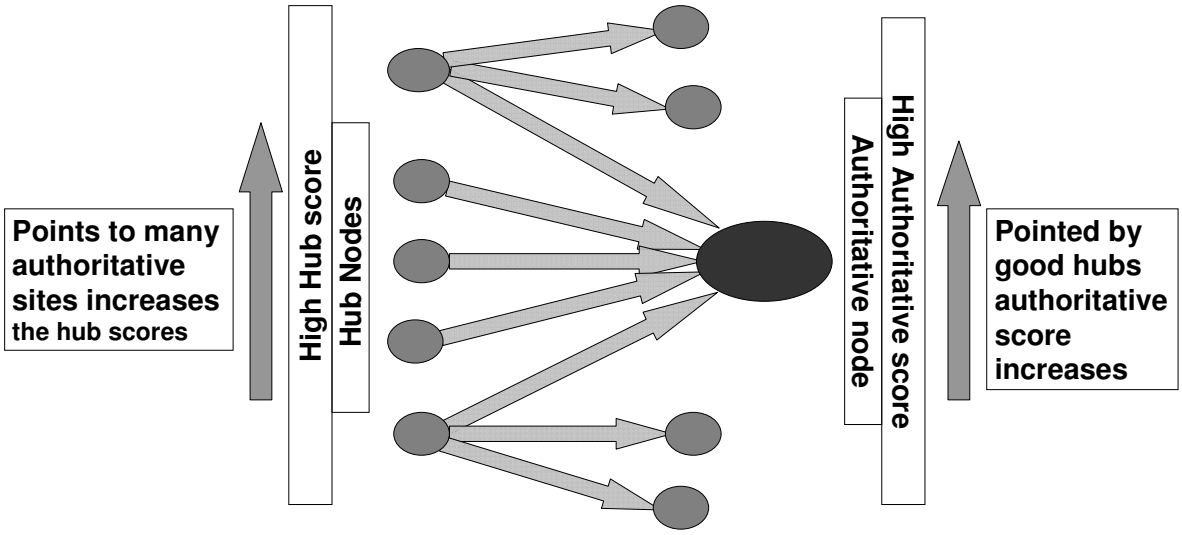

Figure 3: Kleinberg's Authoritative and Hub Nodes

Like 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 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[12].

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. A sample query to prioritize genes associated with cardiomyopathy is shown in Fig-1. However, SPARQL doesn't prioritize the results, we borrowed a technique from M. Sougata et al [12] 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.

3 RESULTS

3.1 Benchmark of the method

To explore the feasibility of our approach in candidate gene prioritization, we randomly selected 40 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 150 genes. The benchmark results were promising, as for 32 out of 40 cases (80%) the related gene is ranked in the top 10 and in 26 cases (65%) ranked in top 5.

3.2 Application

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, a recently reported inherited cardiomyopathy susceptibility region on human chromosome 7 [19]. 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 [29].

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[20]. 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 [30]. K336E mutation in ACTA1(Ranked 2) is associated with fatal hypertrophic cardiomyopathy [30]. A missense mutation of CRYAB (Ranked 5), Arg157His, was

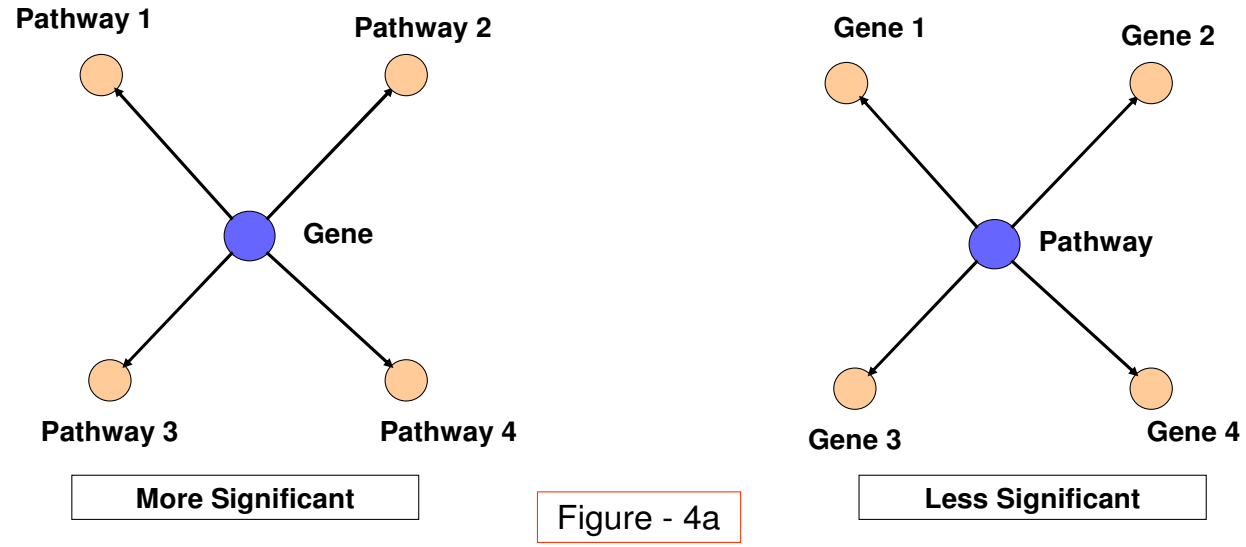

Figure - 4a

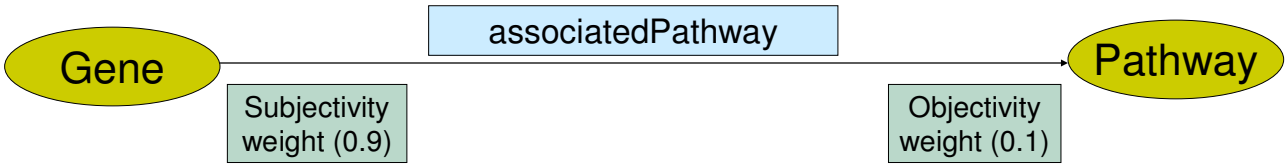

Figure – 4b

Figure 4: 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'*

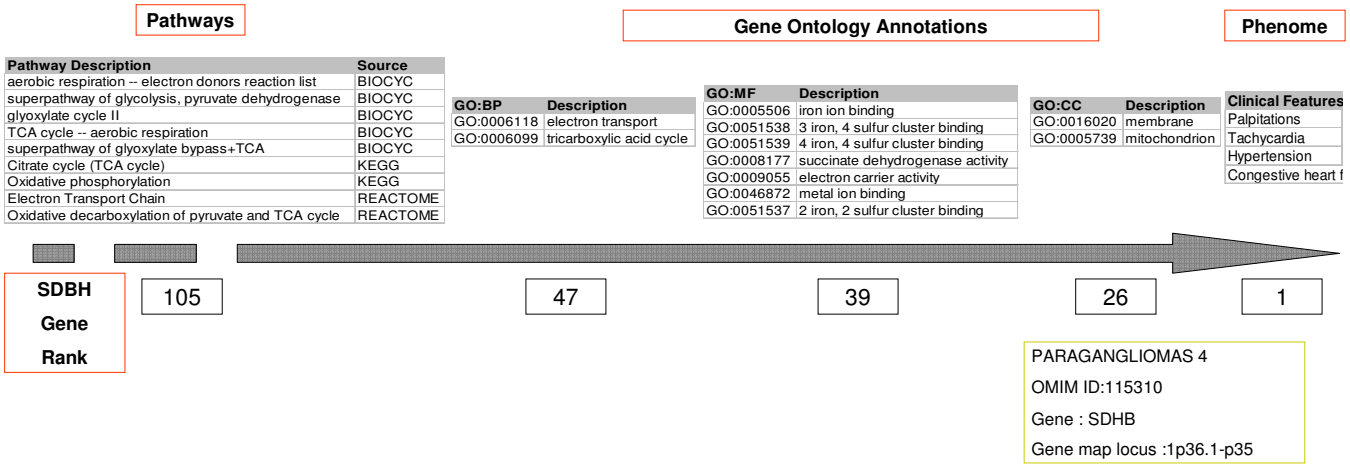

Figure 5: Sequential addition of Genome – Phenome datasets improves SDBH gene ranking implicated in PARAGANGLIOMAS 4

found in a familial DCM patient and the mutation affected the evolutionary conserved amino acid residue among alpha-crystallins [31]. 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 [32]. RYR2 mutations are also known to cause cardiomyopathies and sudden cardiac death [33].

4. ADVANTAGES OF USING SEMANTIC WEB TECHNOLOGIES

4.1 Flexible Integration of Genome to Phenome Networks and Querying

The following example explains how RDF based integrative approaches helped 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-5 explains how flexible integration provided by RDF improves the rank of implicated gene. Using RDF, being a DAG, these layers of information can easily be integrated and mined by graph theoretical algorithms. As a general conclusion, the more relevant data sources we integrated, the better was the overall performance. Moreover, the algorithm requires constant traversals of graph to score each node in the network and SPARQL provides the required graph querying capabilities.

4.2 Adding Context through Semantic Weights

As discussed in the methods section and from Figure-4, by incorporating context specific subjectivity (SW) and objectivity weights (OW), we were able to improve ranking of certain genes. 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.

4.3 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-6 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.

5. 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* [10] and *PROSPECTR* [5] and POCUS [7], 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 and 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 [34].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.

6. 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

```
PREFIX CCHMC:<http://www.cchmc.com/Bio_RDF.owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?pathway
where {
?pathway rdf:type CCHMC:KEGG_Pathway .
}
```

Rank	Pathway	Score
1	Agrin in Postsynaptic Differentiation	0.35737
2	Actions of Nitric Oxide in the Heart	0.27969
3	Stress Induction of HSP Regulation	0.18511
4	Integrin Signaling Pathway	0.185
5	uCalpain and friends in Cell spread	0.185
6	How Progesterone Initiates the Oocyte Maturation	0.1844
7	Signaling of Hepatocyte Growth Factor Receptor	0.15668
8	Y branching of actin filaments	0.15668
9	How does salmonella hijack a cell	0.15668
10	NFAT and Hypertrophy of the heart	0.15668

(a)

```
PREFIX CCHMC:<http://www.cchmc.com/Bio_RDF.owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?pathway
where {
?pathway rdf:type CCHMC:BIOCARTA_Pathway .
}
```

Rank	Pathway	Score
1	Oxidative phosphorylation	3.1938
2	Citrate cycle (TCA cycle)	0.4962
3	Calcium signaling pathway	0.4762
4	Cell Communication	0.3419
5	Tight junction	0.317
6	Focal adhesion	0.3162
7	Leukocyte transendothelial migration	0.2799
8	Regulation of actin cytoskeleton	0.2533
9	Adherens junction	0.2527
10	ATP synthesis	0.2315

(b)

```
PREFIX CCHMC:<http://www.cchmc.com/Bio_RDF.owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?pathway
where {
?pathway rdf:type CCHMC:REACTOME_Pathway .
}
```

Rank	Pathway	Score
1	Electron Transport Chain	1.82998
2	Oxidative decarboxylation of pyruvate	0.45765
3	Gene Expression	0.11796
4	Translation	0.11069
5	Nucleotide metabolism	0.09278
6	Lipid metabolism	0.04762
7	Apoptosis	0.02358
8	Metabolism of sugars	0.01287
9	Xenobiotic metabolism	0.01241
10	Hemostasis	0.01223

(c)

```
PREFIX CCHMC:<http://www.cchmc.com/Bio_RDF.owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?pathway
where {
?pathway rdf:type CCHMC:BIOCYC_Pathway .
}
```

Rank	Pathway	Score
1	aerobic respiration -- electron donors reaction list	0.71909
2	TCA cycle -- aerobic respiration	0.43913
3	glyoxylate cycle II	0.42936
4	superpathway of glycolysis and TCA variant VIII	0.06201
5	TCA cycle variation VIII	0.03
6	gluconeogenesis	0.02748
7	serine-isocitrate lyase pathway	0.01846
8	phenylalanine degradation I	0.01846
9	aspartate degradation II	0.01846
10	glyoxylate cycle	0.01846

(d)

Figure 6: Ranked pathways from various sources from the Bio-RDF associated with differentially expressed genes in human idiopathic dilated cardiomyopathy (DCM) [16]

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.

7. REFERENCES

[1] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion," *Nat Biotechnol,* vol. 24, pp. 537-44, May 2006.

[2] C. Giallourakis, C. Henson, M. Reich, X. Xie, and V. K. Mootha, "Disease gene discovery through integrative genomics," *Annu Rev Genomics Hum Genet,* vol. 6, pp. 381-406, 2005.

[3] S. Rossi, D. Masotti, C. Nardini, E. Bonora, G. Romeo, E. Macii, L. Benini, and S. Volinia, "TOM: a web-based integrated approach for identification of candidate disease genes," *Nucleic Acids Res,* vol. 34, pp. W285-92, Jul 1 2006.

[4] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "SUSPECTS: enabling fast and effective prioritization of positional candidates," *Bioinformatics,* vol. 22, pp. 773-4, Mar 15 2006.

[5] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "Speeding disease gene discovery by sequence-based candidate prioritization," *BMC Bioinformatics,* vol. 6, p. 55, 2005.

[6] J. Freudenberg and P. Propping, "A similarity-based method for genome-wide prediction of disease-relevant human genes," *Bioinformatics,* vol. 18 Suppl 2, pp. S110-5, 2002.

[7] F. S. Turner, D. R. Clutterbuck, and C. A. Semple, "POCUS: mining genomic sequence annotation to predict disease genes," *Genome Biol,* vol. 4, p. R75, 2003.

[8] M. A. van Driel, K. Cuelenaere, P. P. Kemmeren, J. A. Leunissen, and H. G. Brunner, "A new web-based data mining tool for the identification of candidate genes for human genetic disorders," *Eur J Hum Genet,* vol. 11, pp. 57-63, Jan 2003.

[9] N. Tiffin, J. F. Kelso, A. R. Powell, H. Pan, V. B. Bajic, and W. A. Hide, "Integration of text- and data-mining using ontologies successfully selects disease gene candidates," *Nucleic Acids Res,* vol. 33, pp. 1544-52, 2005.

[10] C. Perez-Iratxeta, M. Wjst, P. Bork, and M. A. Andrade, "G2D: a tool for mining genes associated with disease," *BMC Genet,* vol. 6, p. 45, 2005.

[11] J. H. Tim Berners-Lee, Ora Lassila, "The Semantic Web," *Scientific American Magazine,* vol. 284, pp. 29-37, May 2001.

[12] B. Bhuvan and M. Sougata, *Utilizing Resource Importance for Ranking Semantic Web Query Results*, 2005.

[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM,* vol. 46, pp. 604-632, 1999.

[14] B. H. Junker, D. Koschutzki, and F. Schreiber, "Exploration of biological network centralities with CentiBiN," *BMC Bioinformatics,* vol. 7, p. 219, 2006.

[15] S. B. a. L. Page, "The anatomy of a large-scale hypertextual {Web} search engine," *Computer Networks and ISDN Systems,* vol. 30, pp. 107--117, 1998.

[16] K. A. L. irk Koschützki, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl and Oliver Zlotowski, "Centrality Indices," *Lecture Notes in Computer Science,* vol. 34/8, pp. 16-61, 2005.

[17] U. B. a. D. Fleischer, "Centrality Measures Based on Current Flow," *Lecture Notes in Computer Science,* pp. 533-544, 2005.

[18] M. Sougata, B. Bhuvan, and K. Pankaj, "Information Retrieval and Knowledge Discovery Utilizing a BioMedical Patent Semantic Web." vol. 17: IEEE Educational Activities Department, 2005, pp. 1099-1110.

[19] L. Song, S. R. DePalma, M. Kharlap, A. G. Zenovich, A. Cirino, R. Mitchell, B. McDonough, B. J. Maron, C. E. Seidman, J. G. Seidman, and C. Y. Ho, "Novel locus for an inherited cardiomyopathy maps to chromosome 7," *Circulation,* vol. 113, pp. 2186-92, May 9 2006.

[20] R. Grzeskowiak, H. Witt, M. Drungowski, R. Thermann, S. Hennig, A. Perrot, K. J. Osterziel, D. Klingbiel, S. Scheid, R. Spang, H. Lehrach, and P. Ruiz, "Expression profiling of human idiopathic dilated cardiomyopathy," *Cardiovasc Res,* vol. 59, pp. 400-11, Aug 1 2003.

[21] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res,* vol. 32, pp. D258-61, Jan 1 2004.

[22] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Res,* vol. 34, pp. D354-7, Jan 1 2006.

[23] P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas, "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes," *Nucleic Acids Res,* vol. 33, pp. 6083-9, 2005.

[24] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Res,* vol. 33, pp. D428-32, Jan 1 2005.

[25] C. L. Smith, C. A. Goldsmith, and J. T. Eppig, "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information," *Genome Biol,* vol. 6, p. R7, 2005.

[26] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res,* vol. 33, pp. D514-7, Jan 1 2005.

[27] S. Jablonski, "Jablonski's Dictionary of Syndromes & Eponymic Diseases," *Krieger Pub,* vol. 2nd ed, 1991.

[28] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, "Data integration and genomic medicine," *J Biomed Inform,* vol. 40, pp. 5-16, Feb 2007.

[29] E. Bruno, N. Rossi, O. Thuer, R. Cordoba, and L. E. Alday, "Cardiovascular findings, and clinical course, in patients with Williams syndrome," *Cardiol Young,* vol. 13, pp. 532-6, Dec 2003.

[30] J. L. Jefferies, B. W. Eidem, J. W. Belmont, W. J. Craigen, S. M. Ware, S. D. Fernbach, S. R. Neish, E. O. Smith, and J. A. Towbin, "Genetic predictors and remodeling of dilated cardiomyopathy in muscular dystrophy," *Circulation,* vol. 112, pp. 2799-804, Nov 1 2005.

[31] N. Inagaki, T. Hayashi, T. Arimura, Y. Koga, M. Takahashi, H. Shibata, K. Teraoka, T. Chikamori, A. Yamashina, and A. Kimura, "Alpha B-crystallin mutation in dilated cardiomyopathy," *Biochem Biophys Res Commun,* vol. 342, pp. 379-86, Apr 7 2006.

[32] P. J. Kannankeril, B. M. Mitchell, S. A. Goonasekera, M. G. Chelu, W. Zhang, S. Sood, D. L. Kearney, C. I. Danila, M. De Biasi, X. H. Wehrens, R. G. Pautler, D. M. Roden, G. E. Taffet, R. T. Dirksen, M. E. Anderson, and S. L. Hamilton, "Mice with the R176Q cardiac ryanodine receptor mutation exhibit catecholamine-induced ventricular tachycardia and cardiomyopathy," *Proc Natl Acad Sci U S A,* vol. 103, pp. 12179-84, Aug 8 2006.

[33] I. Jona and P. P. Nanasi, "Cardiomyopathies and sudden cardiac death caused by RyR2 mutations: are the channels the beginning and the end?," *Cardiovasc Res,* vol. 71, pp. 416-8, Aug 1 2006.

[34] A. R. Clarke, "Murine genetic models of human disease," *Curr Opin Genet Dev,* vol. 4, pp. 453-60, Jun 1994.