Journal of Biomedical Informatics 41 (2008) 717–729

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge

Ranga Chandra Gudivada [a,c,*], Xiaoyan A. Qu [a,c], Jing Chen [a,c], Anil G. Jegga [b,c], Eric K. Neumann [d], Bruce J. Aronow [a,b,c,*]

[a] Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH 45229-3039, USA
[b] Department of Pediatrics, Cincinnati Childrens Hospital Medical Center, Cincinnati, OH 45229-3039, USA
[c] Division of Biomedical Informatics, Cincinnati Childrens Hospital Medical Center, Cincinnati, OH 45229-3039, USA
[d] Clinical Semantics Group, Lexington, MA 02420, USA

ARTICLE INFO

Article history:
Received 1 September 2007
Available online 23 August 2008

Keywords:
Semantic Web
RDF
OWL
SPARQL
Semantic ranking
Ontologies
Data integration
Bioinformatics
NLP

ABSTRACT

Most common chronic diseases are caused by the interactions of multiple factors including the influences and responses of susceptibility and modifier genes that are themselves subject to etiologic events, interactions, and environmental factors. These entities, interactions, mechanisms, and phenotypic consequences can be richly represented using graph networks with semantically definable nodes and edges. To use this form of knowledge representation for inferring causal relationships, it is critical to leverage pertinent prior knowledge so as to facilitate ranking and probabilistic treatment of candidate etiologic factors. For example, 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 linked to biochemical properties that are shared by other genes 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 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 knowledge bases and can greatly facilitate identification of centrality elements that can lead to specific hypotheses and new insights.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

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 converge on specific genes or features that underlie a disease [1,2]. Quantitative trait loci intervals identified by positional genetics usually include anywhere between 5 and 300 genes [3] and expression studies generate hundreds of unprioritized differentially regulated genes [4]. The identification of the right set of genes from these generated lists for further mutation analysis to associate with the disease under study is termed gene prioritization [5–8]. 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.

The discovery of genes and specific gene variants that cause or modify disease has been shown to be accelerated by knowledge integration and the application of a variety of computational

* Corresponding authors. Address: Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH 45229-3039, USA.
E-mail addresses: gudx6u@cchmc.org (R.C. Gudivada), Bruce.Aronow@cchmc.org (B.J. Aronow).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.07.004

methodologies, in particular to genome-scale experiments [5]. Integrating diverse functional genomic data has several advantages as described by Giallourakis et al. [1]. First, a more comprehensive description of functional gene networks can be formed by essentially combining complementary view-points generated from interrogation of diverse aspects of gene function from different technologies. Second, data integration reduces noise associated with each experimental limitation that limits false positives and increases sensitivity and specificity to detect true functional relationships. 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 [2,3,5–17]. Some of them [4,5,9,10,12] use training gene sets to prioritize candidate test genes based on their similarity with the training properties obtained from the reference set. The significant drawback in these methods is the dependence on there being a sufficiently large number of training set genes. In many practical situations, relevant training sets are not available and results may also vary depending on different approaches used to delineate the particular training set. Though there are methods [2,6–8,11,13,14] that do not require any training set, their potential is limited by their reliance on a small number of data sources. Here, for the first time we utilized Semantic Web (SW) [18] standards and techniques for finding 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 the candidate gene set. The resulting BioRDF (i.e. RDF generated from life science datasets) is a conventional directed acyclic graph (DAG) on to which centrality analysis is applied to score the elements in the network based on their importance within network structure. Centrality analysis determines the relative importance of a node within a graph, by performing a graph theoretic measure on each node [19]. There are several measures to quantify centrality. Here we have utilized *degree centrality* analysis, which considers the number of links incident upon a node. In the context of RDF, resources that have a high in-degree (the number of links coming into a node in a directed graph) or out-degree (the number of links going out of a node in a directed graph) implicate a highly significant node. Central elements in biological networks are generally found to be essential for viability and their delineation within a network leads to new insights and potential to generate new hypotheses [20]. In this approach, score of each gene depends on the functional importance inferred from the genomic knowledge combined with the clinical features representing phenomic knowledge. Centrality measures are calculated from a modified version [21] of the *Kleinberg algorithm* [22] similar to Google's Page rank algorithm [23] extended for the Semantic Web. While Semantic Web querying languages do not per se naturally rank the retrieved results from RDF graphs, we have adapted a technique described by M. Sougata et al. [21,24] for domain-specific ranking to rank the retrieved genes from BioRDF using SPARQL (http://www.w3.org/TR/rdf-sparql-query/). RDF graphs provide the ability to aggregate and recombine loosely associated disease and molecular information into a formal knowledge structure. This semantic mashup can be viewed together or analyzed as a complete set. In addition, semantic mashup are not just for viewing facts, they can support analytical lenses and algorithms for uncovering deeper meaningful associations.

Thus, although there have been several other approaches developed that either include purely genomic data [3,5–7,10,25] or genomic data combined with either human [2,8,9,11,12,14,26] or mouse phenomic [4] data sets in order to expedite disease gene search, our approach enables for the first time systematic gene prioritization without the assertion of a focus training set by utilizing both mouse phenotypes and human disease clinical features as well as their GO and pathways relationships. Our method does not use any training data set, but extends the earlier hypothesis that majority of the disease-causal genes are functionally important and share clinical features with related diseases [5,8,11,12]. We reasoned that an integrative genomic–phenomic approach utilizing the available human gene annotations including human and mouse phenomic knowledge will provide more comprehensive and valid disease candidate gene identification and prioritization. In this study, we have focused on cardiovascular system diseases (CVD). We tested our hypothesis by prioritizing genes from the recently reported (a) hypertrophic cardiomyopathy susceptibility loci (chromosome 7p12.1–7q21) [27] (b) dilated cardiomyopathy loci (chromosome 10q25–26) [28] and (c) among genes differentially expressed in dilated cardiomyopathy [29].

2. Methods

2.1. Knowledge sources

Genomic and phenomic knowledge representation was accomplished by RDF conversion of datasets from multiple data sources (see Fig. 1). These are described as follows:

2.1.1. Genomic knowledge sources

(1) Gene Ontology (GO) [30] was downloaded from Gene Ontology 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 [31], BioCarta (http://www.biocarta.com/), BioCyc [32], and Reactome [33]. 4772 human genes had at least one pathway association (a total of 672 pathways).

2.1.2. Phenomic knowledge sources

(1) Mammalian Phenotype (MP) ontology [34], 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) A total of 977 records (423 have at least one implicated gene) were downloaded in XML format from OMIM [35] by searching for terms "cardiovascular" or "heart" or "cardiac" occurring in clinical synopsis (CS) 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 also parsed each TX section of OMIM record as it provides additional clinical features to the ones available from CS section, which is evident from Fig. 2. 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).

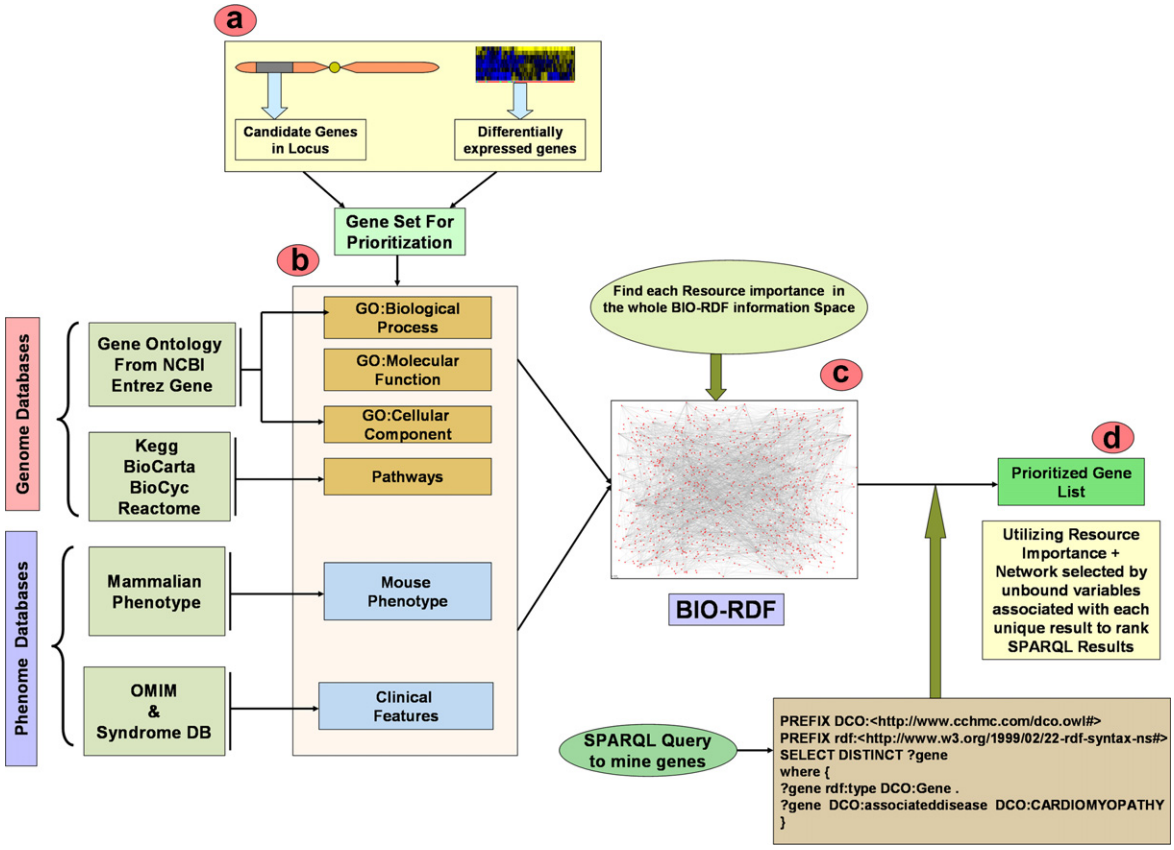

Fig. 1. Schema diagram. (a) Test gene set is obtained from a locus identified by linkage analysis or differentially expressed genes from a microarray experiment. (b) Genome and phenome knowledge sources considered to create BioRDF includes GO: Molecular Function, GO: Biological Process, GO: Cellular Component, Pathways, Mammalian Phenotype, OMIM and Syndrome DB. (c) Each resource in the BioRDF (information space) 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.

(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 their website (http://www.nlm.nih.gov/archive//20061212/mesh/jablonski/syndrome_toc/toc_c.html). This resource was developed by Stanley Jablonski [36] and consists of structured descriptions of approximately 1600–2000 syndromes belonging to 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. 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 were parsed using JAVA XML scripts from the downloaded XML files. The CS section of OMIM and the 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 [37,38]. It is 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. Fig. 3 provides an example of overcoming orthographic (spelling variants) problem inherent among clinical terms represented in OMIM records by mapping to UMLS concepts. We used an online version of MetaMap program (SKR-MetaMap) that is available as part of Semantic Knowledge Representation project (SKR) (http://skr.nlm.nih.gov/), that aims to provide a framework for exploiting UMLS knowledge resources for NLP.

The extracted clinical features were uploaded into the SKR-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* (CUIs) and the *Semantic Type* it belongs to from the list of final candidate mappings. To avoid the 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 by selecting scores ranging from 570 to 1000 and after careful manual curation incorrectly assigned concepts were removed.

The online SKR-MetaMap works well for short phrases but requires exceptionally long processing times when 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) [39], 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

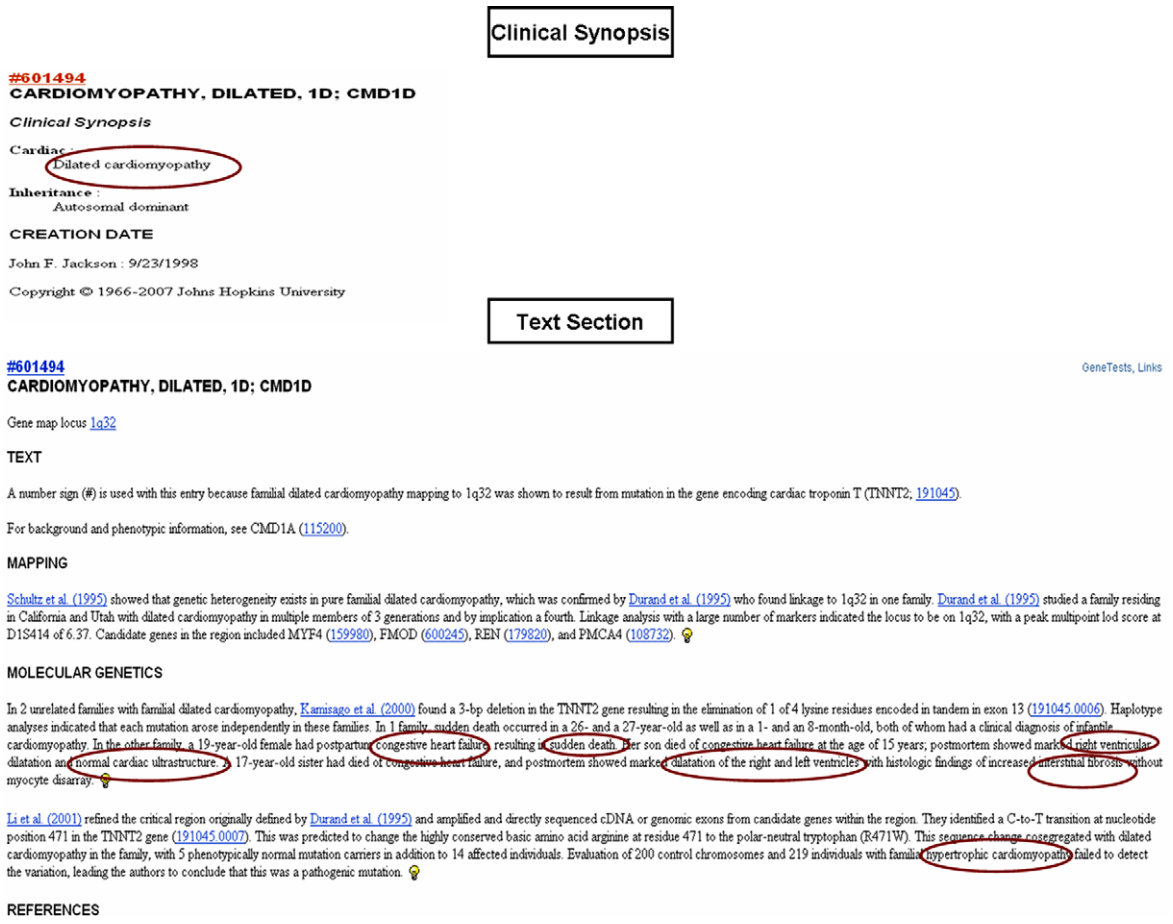

Fig. 2. Text (TX) and clinical synopsis (CS) sections from OMIM for cardiomyopathy dilated 1D (OMIM No. 601494). As shown, TX section provides additional clinical features to the ones mentioned in CS.

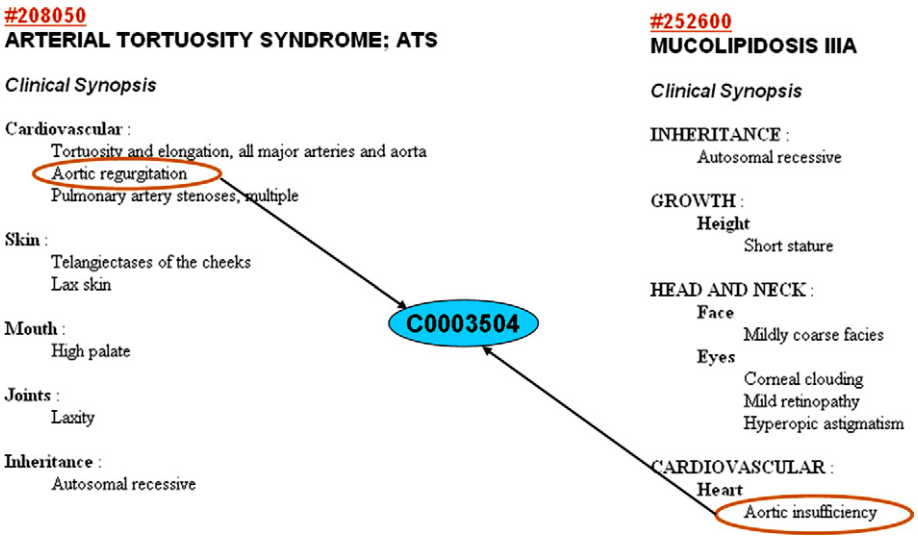

Fig. 3. clinical synopsis sections for arterial tortuosity syndrome (OMIM No. 208050) and Mucolipidosis IIIA (OMIM No. 252600) disorders. Figure explains overcoming orthography by semantic normalization of clinical features to UMLS concepts.

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*' semantic group. For each concept belonging to this group, preferred names and synonyms were extracted and supplied to the gazetteers. 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 1 provides the statistics before and after performing semantic normalization of the OMIM clinical features to UMLS concepts and the table does not indicate the MetaMap performance evaluation. MetaMap mitigates the manual curation to a large extent by controlling semantic types and keeping us focused on less scored mappings and this process is also scalable to a much large data sets. The advantage of using UMLS concepts instead of raw clinical features from unstructured text extremely reduced the total clinical feature space by around 50%. We have to consider the entire UMLS for the semantic normalization of OMIM clinical features as no single terminology or ontology is sufficient to provide the necessary coverage (Table 2A and B).

2.3. Mapping clinical features to genes

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

2.4. Generating RDF

The Resource Description Framework (RDF), an official W3C recommendation, provides a generic framework to describe entity properties, relationships, and constraints, and can be used to form directed acyclic graph (DAG) representations of multidimensional data and web resources. It is a semi-structured data model in which complex relations can be readily modeled [40]. 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'>, expresses 'ATM' as subject, 'is a' the property and 'Gene' as object of the statement. Disease Card Ontology (DCO), an ontology currently under internal development [41] 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 [42,43] 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 from local relational databases to create BioRDF instantly on the fly for the specific disease and gene set under study. 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 (Fig. 1). Fig. 4 provides a portion of DCO and associated BioRDF. As we are focusing on CVD, mouse phenotypes are restricted under '*cardiovascular system phenotype*' a parent node in the Mouse Phenotype Ontology.

2.5. Ranking on Semantic Web (SW)

In real world Semantic Web, most queries will result in large numbers of retrieved results. Therefore, developing efficient information retrieval techniques for discovering relevant knowledge will be crucial towards realizing the vision of Semantic Web. In our approach, we see the ranking of retrieved disease genes as essential since researchers will tend to consider only the first few results. Our approach to ranking Semantic Web resources is based on an algorithm developed and successfully implemented in the BioPatentMiner System [24] that itself was an extension to an

Table 1
Total clinical features extracted from OMIM and Syndrome DB before and after performing semantic normalization (mapping to UMLS concepts)

Clinical features	Total extracted features	Total features after semantic normalization	% Of clinical feature reduction after semantic normalization
Clinical symptoms	16979	8504	50.08
Affected anatomy	8062	3364	41.7

Table 2

Source name	Source abbreviation	Version	Number of mapped concepts
(A) List of the top 10 source terminologies in UMLS to which OMIM clinical features have associated concepts			
SNOMED clinical terms	SNOMEDCT	2007	6587
Medical dictionary for regulatory activities terminology (MedDRA)	MDR	10	4391
ICPC2–ICD10 thesaurus	ICPC2ICD10ENG	200412	3467
UMLS Metathesaurus	MTH	2007	3052
Medical subject headings	MSH	2007	3038
Online mendelian inheritance in man	OMIM	2007	2993
National cancer institute thesaurus	NCI	2006	2223
Canonical clinical problem statement system	CCPSS	1999	2046
National drug file—reference terminology	NDFRT	2004	2021
Dxplain	DXP	1994	1978
			Number of clinical features missing in SNOMED (2007)
(B) Illustrating the need to map to entire UMLS to achieve better coverage rather than restricting to one single terminology such as SNOMED			
Medical dictionary for regulatory activities terminology (MedDRA)			599
ICPC2–ICD10 thesaurus			314
UMLS Metathesaurus			285
Medical subject headings			305
Online mendelian inheritance in Man			313
National cancer institute thesaurus			397
Canonical clinical problem statement system			160
National drug File—reference terminology			96
DXplain			141

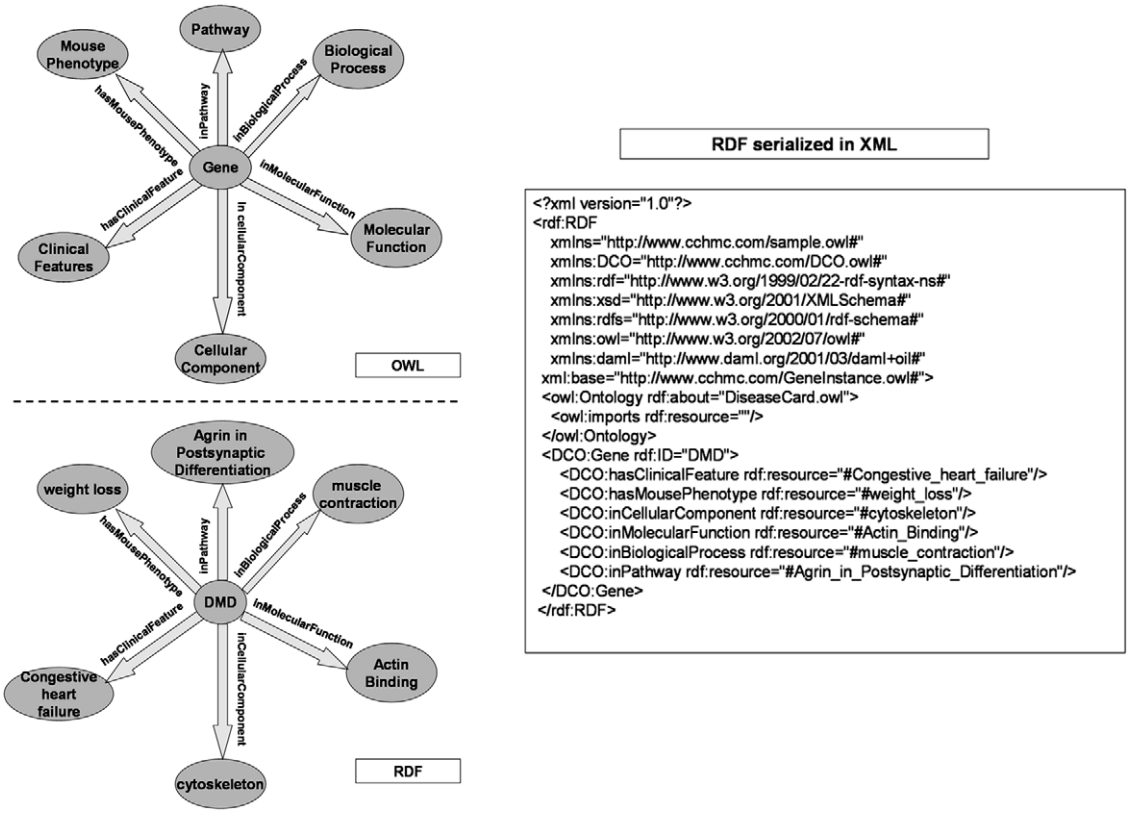

Fig. 4. Portion of BioRDF generated for DMD gene based on the DCO ontology. The upper network is the ontology providing the required semantics (Classes) for the lower RDF network consisting instance data.

earlier WWW link-analysis algorithm [22] to identify relevant web pages based on the number of pages linking to it and also the importance of linking pages. The extended algorithm considers specific aspects of Semantic Web such as information complexity compared to a traditional web since it contains different kinds of resources and relations between them as well as ontological relations and references. In addition, it follows the same principle as Google that ranking the search results should not be determined just for specific queries but rather by the importance of the results in the overall information space (RDF graph) [21]. Google search result ranking relies on web content analysis performed over the full information space prior to any query and the same logic can also be applied in querying for the disease genes on an integrative functional Bio-RDF network created for a particular disease. The algorithm is recursive and the score of each node is passed to the adjacent node in the next iteration, until score becomes constant with further iterations. This score indicates the relevance of the node in the network based on the importance it has in relative to the overall disease information space. In the next section, we briefly describe the algorithm and metrics in calculating scores for each resource. For a more complete in-depth analysis and explanation of the algorithm, refer to the original paper [24,44].

2.5.1. Calculating resource importance

In the world of Semantic Web, a resource can be considered relevant if it has relations with many other resources where the meaning and significance of these other resources have been recursively defined as relevant with respect to their associated resources. Resource relevance, scoring RDF network elements according to their idiosyncratic defining relationships within the network structure, can be calculated from the complete set of these relationships within the RDF graph set. In the context of a graph, resources that have a high in-degree or out-degree should be considered relevant, i.e. may contain causal or predictive (correlative) relations. In SW networks (graphs), two important metrics were defined to estimate the importance of each resource, *Subjectivity Score* (*SS*) and *Objectivity Score* (*OS*) parallel to Kleinberg's [22] *hub* and *authority scores* for the WWW graph (Fig. 5). Kleinberg not only considers the number of links to and from a node but also the relevance of linked nodes. Accordingly, if a resource in SW is pointed to by a resource with high SS, its OS increases. Conversely, if a resource points to a resource with a high OS, its SS is increased. Initially these scores are set to 1.0 and resources with high subjectivity/objectivity scores are the subject/object of many of RDF triples.

2.5.2. 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 '*associatedPathway*' where it links a gene to a pathway it has role in. A gene associated with multiple pathways can be considered to be more relevant than compared to a pathway having multiple genes because any mutation in the multipathway-linked gene could affect several pathways manifesting into a disease. Therefore, importance of a pathway resource should not increase if it has many genes. However, the relevance of a gene resource obviously depends on the pathways it's associated with, paralleling the causal flow from gene products to the pathways in which they participate. Fig. 6 illustrates the significance of semantic weights on gene-pathway association. In order to influence the scoring scheme, each property is assigned with initial zsubjectivity and objectivity weights, which control the sub-

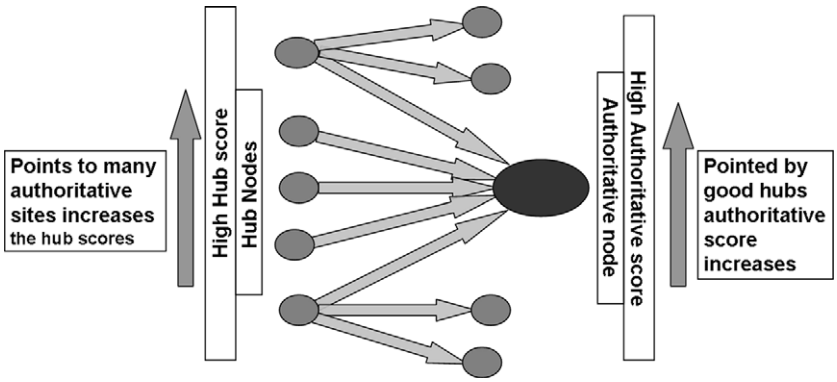

Fig. 5. Kleinberg's authoritative and hub nodes.

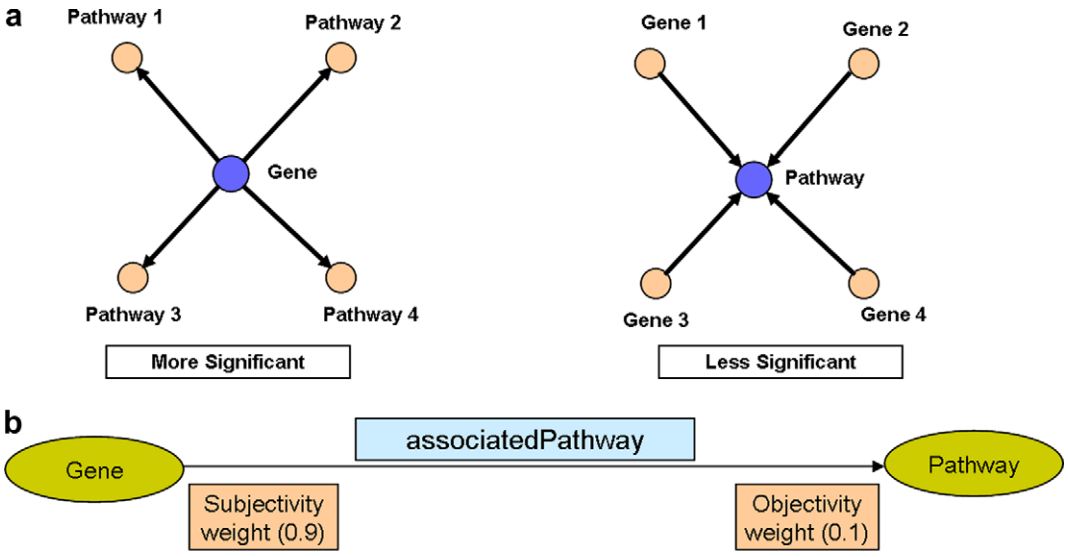

Fig. 6. (a) Illustrating that the significance of a gene associated with multiple pathways is considered more important compared to a pathway having multiple genes within a disease context. (b) Assigning subjectivity and objectivity weights to the property '*associatedPathway*' for the triple '*gene—associatedPathway—Pathway*'.

ject/object scores (resource importance) for that property. Consequently, properties like '*associatedPathway*' *are* assigned with higher subjectivity weight and lower objectivity weight. As gene is the subject of all triples (from Fig. 4), every property is assigned with higher subjectivity weights and lower objectivity weights. Since the relative strength between subjectivity and objectivity is really what important, the choice of exact weights can be arbitrary, however, one constraint is that for each property the sum of subjectivity and objectivity weights must be equal to 1.0. We have chosen a subjectivity weight of 0.9 and objectivity weight of 0.1. The modified Kleinberg's algorithm [21] to calculate Subjectivity and Objectivity scores of Semantic Web resources with predefined Subjectivity and Objectivity weights is as follows:

1. Let R be the set of resources (nodes) and E be the set of properties (edges) in the BioRDF graph.
2. For every resource r in R, let $S[r]$ be its subjectivity score and $O[r]$ be its objectivity score.
3. Initialize $S[r]$ and $O[r]$ to 1 for all r in R.
4. While the vectors S and O have not converged:
   (a) For all r in R, $O[r] = \sum_{(r1,r)\in E} S[r1]^{*}\text{objWt}(e)$ where objWt is the objectivity weight of the property representing the edge.
   (b) For all r in R, $S[r] = \sum_{(r,r1)\in E} O[r1]^{*}\text{subWt}(e)$ where subWt is the subjectivity weight of the property representing the edge.
   (c) Normalize the S and O *vectors.*

The modification is that while determining the subjectivity and objectivity scores of a node we multiply the scores of the adjacent vertex by the subjectivity and objectivity weights of the corresponding link. This will ensure that the scores of certain resources are not influenced by the total number of resources it's associated with for a particular property. For example, a low objectivity weight for the '*associatedPathway*' property will ensure that the objectivity scores of pathway resources are not increased by the number of genes that pathway is associated with. As with the original Kleinberg algorithm, our modified version also terminates with all the vectors converging for any Semantic Web graph. Convergence is defined when the subjectivity and objectivity scores for all resources become stable after finite iterations, at which the program is automatically terminated. Finally, the importance of each resource $I[r]$ is determined by adding its corresponding subjectivity and objectivity score as follows: $I[r] = S[r] + O[r]$.

2.5.3. Ranking the retrieved results

Search result ranking is an important research topic in information retrieval. The node scores used for ordering the results are not determined by a specific query but calculated prior through the relevance of the data nodes in the overall information space. But for every issued query, the resultant ranked list of nodes are

identified by the SPARQL-SELECT clause and sorted according to their pre-calculated relevance scores. We used ARQ (http://jena.sourceforge.net/ARQ/), a query engine for Jena that supports SPARQL, a RDF query language. A sample query to prioritize genes associated with cardiomyopathy is shown in Fig. 1. However, SPARQL does not in itself prioritize the results, hence we borrowed a technique from Bhuvan and Sougata [21] which adds an extra computational layer to rank the retrieved 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 retrieve the associated graph for each result using 'CONSTRUCT' query form of SPARQL [45], and compute a score for every result. The original equation was designed to handle queries ranging from simple to complex and calculated a score for the relevance of each result by using various parameters associated with it. But since as we are prioritizing only genes, the query is more focused and assumes that if a gene has high relevance in the overall semantic graph, their ranking should be correspondingly higher. Therefore, most of the variables in the original equation are assigned 0, but could be incorporated to handle complex queries such as prioritizing genes associated with a particular pathway while also being linked to a specific high-scored Gene Ontology class.

3. Results

3.1. 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 at least one implicated gene with associated clinical synopsis. The algorithm was not provided with any explicit link between target gene and the disease to validate that our method detects the true functional relationship between the disease and the gene. For every OMIM disease from our dataset, we extracted the genes from the locus specified in the OMIM database. On an average we ensured that each list contained around 300 genes including the implicated gene. These gene lists are used to validate how efficient our approach can be in finding the real implicated gene from the other non-disease genes (~300 genes) in that specific locus. The benchmark results were quite promising, since in 44 out of 60 cases (74%) the related gene is ranked in the top 10 and in 33 cases (55%) ranked in top 5.

3.2. Application

We tested the efficacy of our method in prioritizing candidate genes from cardiovascular disease (CVD)-implicated genomic regions (from LOD scores) and from differentially expressed genes from expression studies.

3.2.1. Prioritizing candidate genes from CVD-implicated genomic regions

Linkage analysis is a proven method to associate diseases with specific genomic regions. However, these regions are often large, containing hundreds of genes, which make experimental or automated methods employed to identify the correct disease gene difficult and costly. We used our integrative based ranking approach to prioritize candidate genes from the CVD-implicated genomic regions. As test sets, we used known gene lists from 2 loci recently implicated in cardiomyopathy [27,28].

3.2.1.1. 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 susceptibility region for inherited cardiomyopathy on human chromosome 7 [27]. Mutations in the top ranked genes (Table 3), namely, GTF2IRD1 [46–48], GTF2I [49,50], ELN [51–53], LIMK1 [54–56], and BAZ1B [57–59] (in mouse or human or both) have been associated with Williams–Beuren Syndrome(OMIM ID: 194050). Though this syndrome is primarily recognized as a mental retardation syndrome, it is also associated with cardiovascular symptoms such as atrial septal defect, supravalvar aortic stenosis and less frequently hypertrophic cardiomyopathy [60].

Table 3
Prioritized genes from loci 7p12.1–7q21 associated with hypertrophic cardiomyopathy on chromosome 7

Rank	Gene symbol	Score
1	GTF2IRD1	173.334
2	GTF2I	120.1975
3	ELN	93.53132
4	SBDS	77.47414
5	EGFR	42.30714
6	LIMK1	40.77264
7	YWHAG	38.65037
8	BAZ1B	20.52658
9	ZNF117	9.506009
10	ZNF273	8.496438

3.2.1.2. Prioritization of genes at a locus for dilated cardiomyopathy on chromosome 10q25–26. After prioritizing the 68 genes in the chromosome 10q25–26 region (~9.5 mega bases, locus for cardiomyopathy, diffuse myocardial fibrosis, and sudden death) [28], we identified FGFR2 as the top ranked gene. FGF signaling via FGFR2 regulates myocardial proliferation during midgestation heart development and in the absence of this signal newborn mice develop dilated cardiomyopathy [61]. From a study [62], comparing the GRK5 (second ranked) expression in patients with left ventricular volume-overload disorders and dilated cardiomyopathic hearts, a relation exists between the expression of GRK5 and alterations in myocardial β-adrenoceptor signaling in volume-overload. The result point to myocardial GRK5 regulation in cardiac disease localized to ventricles. Jahns et al. [63] have provided direct evidence that an autoimmune attack directed against the cardiac β(1)-adrenergic receptor, ADRB1 (third ranked) may play a causal role in dilated cardiomyopathy (DCM). A recent study reports the use of ADRB1 as a prognostic marker, a risk predictor, and adverse clinical effects by stimulating antiβ1-antibodies in DCM [64]. Table 4 provides the top 10 prioritized genes at this loci for dilated cardiomyopathy.

3.2.2. Prioritizing Candidate Genes from the Differentially Expressed genes in CVD

Microarray analysis is a powerful technique for high-throughput, global transcriptonomic profiling of gene expression. It holds great promise for analyzing the genetic and molecular basis of var-

Table 4
Prioritized genes from loci 10q25–26 associated with dilated cardiomyopathy on chromosome 10

Rank	Gene symbol	Score
1	FGFR2	144.9478
2	GRK5	138.7307
3	ADRB1	122.455
4	TIAL1	100.2871
5	EMX2	97.13126
6	GFRA1	82.81983
7	BUB3	47.78852
8	DMBT1	46.70792
9	SLC18A2	20.62583
10	PRLHR	19.57693

ious complex diseases and permitting the analysis of thousands of genes simultaneously, both in diseased and non-diseased tissues and/or cell lines [65]. However, it often provides researchers with too many candidates without necessarily identifying causative elements. To assess our prioritization approach with such studies, we used a dataset of differentially expressed genes in human idiopathic dilated cardiomyopathy [29].

3.2.2.1. Gene prioritization of differentially expressed genes in human idiopathic dilated cardiomyopathy (DCM). We used our prioritization approach to rank 216 differentially expressed genes (Table 5 lists the top 10 genes) from the expression profiles of myocardiac biopsies from 10 DCM patients [29]. The top ranked gene is DMD, which is already well known in cardiac function and malformation. Specific DMD gene mutations may protect against or inhibit development of DCM. The K336E mutation in ACTA1 (Ranked 2) is associated with fatal hypertrophic cardiomyopathy [66]. 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 α-crystallins [67]. Although GJA1 (ranked 8th) is not associated with hypertrophic cardiomyopathy, but disturbances in Cx43 expression and localization are reported to influence heart embryogenesis and maturation and also contribute to hypertrophy and dysfunction of the right ventricle, including arrhythmias in children with tetralogy of fallot [68]. RYR2, ranked 10th in our list, encodes ryanodine receptor found in the cardiac muscle sarcoplasmic reticulum. Mice with the R176Q cardiac RYR2 mutation exhibit catecholamine-induced ventricular tachycardia and cardiomyopathy [69]. RYR2 mutations are also known to cause cardiomyopathies and sudden cardiac death [70].

4. Advantages of using Semantic Web technologies

4.1. Flexible integration and query of genomic and phenomic networks Querying

An important feature of our framework is the ability to include multiple knowledge sources related to different disease features for modeling and prioritization. RDF provides a very flexible way to integrate different layers of information and also to mine the integrated network by applying graph theory-based analytical algorithms. We assessed whether our sequential integrative genomic–phenomic approach is capable of prioritizing implicated genes for the 60 sample diseases. Sensitivity and specificity values are computed for the 60 prioritizations using the methodology described in [4,5]. Sensitivity refers to the frequency (% of all prioritizations) of all known disease implicated genes that are ranked above a particular threshold position. Specificity refers to the percentage of actual non-implicated genes ranked below this threshold which is different from negative predictive value which states that the proportion of less ranked genes that are truly non-implicative. We plotted rank receiver operating characteristic (ROC) curves to prove that increasing the number of heterogeneous knowledge bases enhances the probability in predicting the disease implicated gene. ROC curves from Fig. 7 illustrate that sequential addition of genomic to phenomic knowledge integration improves the overall performance of ranking. The greater the area under the curve (AUC) the better the performance and as can be seen from Fig. 7, the area with all the sources is comparatively larger than all the other areas with partial sources, thus supporting our hypothesis.

In addition to ROC curves, the following example illustrates how RDF based integrative approaches assisted to home in on the gene SDHB underlying Paragangliomas 4 (OMIM ID: 115310), a disorder having several cardiovascular symptoms (palpitations, tachycardia, and hypertension). SDHB is one of the 245 genes located at the genomic region 1p36.1–p35. Fig. 8 explains how flexible and incremental integration provided by RDF improves the rank of the implicated gene. To conclude, RDF facilitates the flexible and modular additions of specific knowledge sources to enhance its overall performance. Moreover, the algorithm also requires repeated traversals of the graph with each database addition to properly 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 also from Fig. 6, incorporating context specific subjectivity (SW) and objectivity weights (OW) improved ranking of certain genes. We generated ROC curves with and without semantic weights (Fig. 9) by including all knowledge sources. Fig. 9 clearly illustrates improved overall performance in ranking by assigning weights to properties but as the change in the ranking are only for few genes we did not consider doing a further statistical test of these two ROC curves. For example, the ACADVL gene implicated in mitochondrial very-long-chain acyl-CoA dehydrogenase deficiency (as evidenced in OMIM ID: 201475) ranked 53 without any Subjectivity and Objectivity weights, but improved to rank 9 after including weight functions.

4.3. Ability to investigate other resources (apart from genes) in BioRDF

As every resource is scored in the integrated BioRDF information space, we can issue further SPARQL queries to retrieve and prioritize other entities (apart from genes), such as pathways. Using the Human Idiopathic DCM example (Section 3.2.2.1), we queried further for the important pathways. This provides evidence of other relevant entities shared in the network to corroborate our initial findings. Fig. 10 illustrates the resulting pathways and SPARQL queries retrieved from multiple sources. This feature is particularly useful in expression studies as the differentially expressed genes are already related in a particular disease context.

Table 5
Prioritized genes from differentially expressed genes in human idiopathic dilated cardiomyopathy

Rank	Gene symbol	Score
1	DMD	91.29173
2	ACTA1	64.82657
3	UQCRB	53.62478
4	SDHB	50.8915
5	CRYAB	46.62995
6	SDHA	40.74193
7	LDB3	40.38424
8	GJA1	37.38691
9	ACTC1	34.42755
10	RYR2	20.97512

5. Discussion

Our approach to the prioritization of candidate genes differs from other methods in multiple ways, beginning with more extensive coverage of knowledge bases, flexible data integration methods, and the application of novel mining algorithms. To the best of our knowledge, apart from *G2D* [8], *PROSPECTR* [7], and POCUS [6], most of the current tools for candidate gene prioritization use training gene sets. But in many cases, training gene sets are not available and results are highly dependent on the quality and relevance of the training set used. G2D uses MeSH (www.nlm.nih.gov/mesh) disease terms from publications as clinical fea-

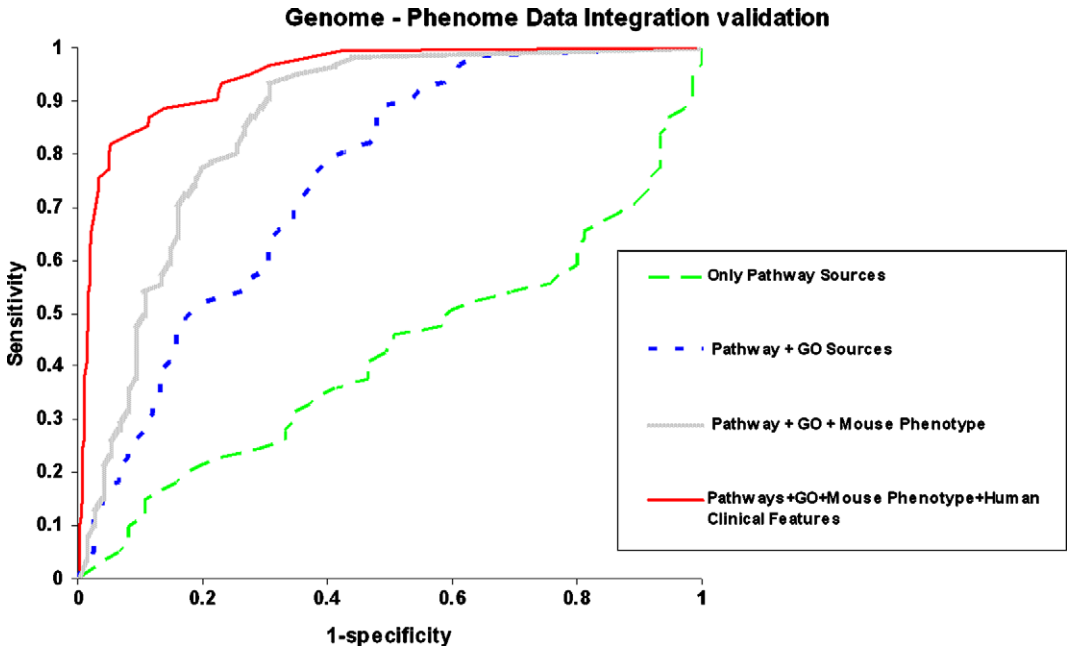

Fig. 7. Rank ROC curves for validating sequential integrative approach in prioritizing the implicated gene (out of 300 genes on average) from the loci associated with the 60 sample OMIM diseases. The 4 curves, represented in different colors are associated with sequential integration of different genomic–phenomic knowledge sources. The data sources associated with each ROC curve are indicated on the figure.

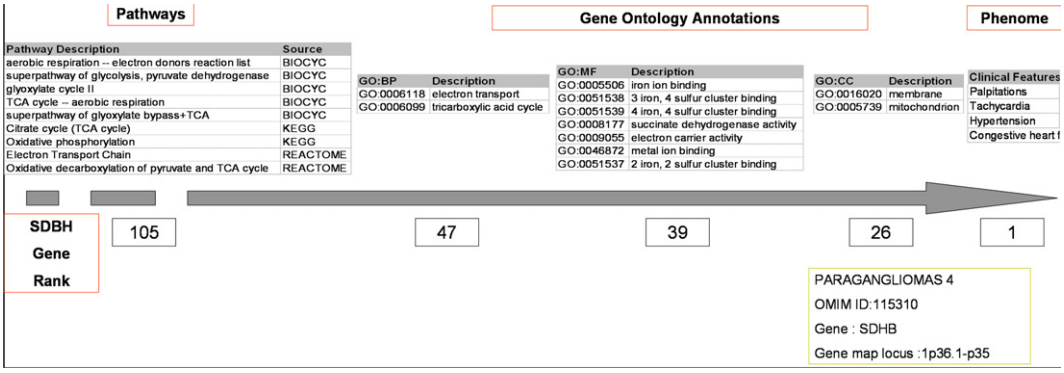

Fig. 8. Sequential addition of genome–phenome knowledge improves ranking of SDBH gene implicated in PARAGANGLIOMAS 4.

tures associated with each OMIM disease. 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 these other approaches integrate both human and mouse clinical features although the mouse is the key model organism for the analysis of mammalian developmental, physiological, and disease processes [71].Our method has two phases, first to find the biologically functional important genes from the test set by integrating multiple genomic knowledge sets. This relevance is scored from their participation in multiple pathways, biological processes and molecular functions independent of any particular disease. In the next phase, we apply specific disease context to the genomic network by adding phenotypic or clinical features relevant to the disease under study (Ex: All clinical features associated with the test genes restricted to CVD in OMIM). This step improves the ranking of those specific genes, considered important from relevant genomic knowledge and also associated with clinical features related to the studied disease. In general, we are applying network centrality analysis to rank resources according to their relevance within the BioRDF graph. Moreover, here, the relevance of a resource is properly enhanced by integrating multiple diverse knowledge sources (from genome to phenome) into the RDF information space. It is also evident from the earlier exemplary work [26,38,72–75] that integration and mining of phenomic and genomic knowledge provides deep insight in elucidating disease–molecular relationships. Additionally, resource ranking is performed semantically by including contextual semantic weights on the properties connecting the resources, which effectively insert general causal relations (such as genes influencing pathway behavior) into the prioritization process. 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. However, as more quality data becomes available and is integrated into BioRDF, we believe the errors will be washed out. At present, BioRDF graphs are generated instantly by retrieving knowledge from local relational databases but the future versions will access a native RDF triple store to extract large subsets of graphs for a particular disease and gene set. We are also planning to move towards using a locally installed version of MetaMap as it can easily handle large sections of free text [38,76] in contrast to the online version.

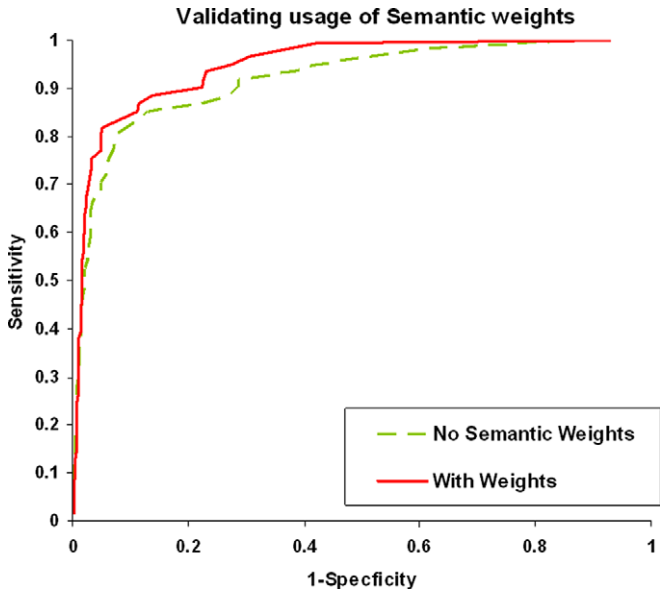

Fig. 9. Rank ROC curves for validating the improvement in overall performance in ranking the implicated gene (out of 300 genes on average) including all sources with and without semantic weights.

6. Conclusion

We have used for the first time in human disease gene prioritization a combination of both mouse phenotype and human disease clinical features from OMIM clinical synopsis. In addition to such extended coverage of knowledge sources used, we have also shown for the first time that one can leverage Semantic Web standards and techniques applied to a specific biological problem. The direct use of W3C's RDF and OWL standards for knowledge integration, the application of network centrality analysis for mining and the retrieving of ranked results using graph query languages such as SPARQL. Although in this current study we focused on the cardiovascular system, our approach can be applied to any group of genes or diseases. One immediate application could be to apply our methods to all OMIM diseases (around 1554) having known loci but unknown molecular basis. As the functional annotations of human and mouse genes improve over time we envisage a proportional increase in the performance and robustness 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 from any source including its homologs and their phenotypes. Consequently, this will enable targeted research on the contribution of genetic mutations towards diseases that will provide specific leads towards novel diagnostic

```
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

Fig. 10. Ranked pathways from various sources of the BioRDF graph associated with differentially expressed genes in human idiopathic dilated cardiomyopathy (DCM) [16].

and therapeutic approaches. Our objective in this manuscript is to make a compelling case and provide evidence that no other single open standard exists for mapping any data record entry from any DB into a common graph space and along with associated ontologies/semantics. Graph models of data are essential to apply algorithms such as Page rank and SW allows Page rank to apply not to just pages but to individual data objects.

Acknowledgment

This paper is an extension of our workshop paper 'A Genome–Phenome Integrated Approach for Mining Disease-Causal Genes using Semantic Web' published in WWW2007/Health Care and Life Sciences Data Integration for the Semantic Web (http://www2007.org/workshop-W2.php). Supported by NCI UO1 CA84291-07 (Mouse Models of Human Cancer) and the State of Ohio Third Frontier Wright Center of Innovation Center for Computational Medicine.

References

[1] Giallourakis C et al. Disease gene discovery through integrative genomics. Annu Rev Genomics Hum Genet 2005;6:381–406.
[2] Tiffin N et al. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 2005;33(5):1544–52.
[3] van Driel MA et al. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. Eur J Hum Genet 2003;11(1):57–63.
[4] Chen J et al. Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics 2007;8(1):392.
[5] Aerts S et al. Gene prioritization through genomic data fusion. Nat Biotechnol 2006;24(5):537–44.
[6] Turner FS, Clutterbuck DR, Semple CA. POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol 2003;4(11):R75.
[7] Adie EA et al. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics 2005;6:55.
[8] Perez-Iratxeta C et al. G2D: a tool for mining genes associated with disease. BMC Genet 2005;6:45.
[9] Rossi S et al. TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res 2006;34(Web Server issue):W285–92.
[10] Adie EA et al. SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics 2006;22(6):773–4.
[11] Freudenberg J, Propping P. A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 2002;18(Suppl. 2):S110–5.
[12] Lage K et al. A human phenome–interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 2007;25(3):309–16.
[13] Masseroli M, Galati O, Pinciroli F. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Res 2005;33(Web Server issue):W717–23.
[14] Gaulton KJ, Mohlke KL, Vision TJ. A computational system to select candidate genes for complex human traits. Bioinformatics 2007;23(9):1132–40.
[15] English SB, Butte AJ. Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes. Bioinformatics 2007;23(21):2910–7.
[16] Calvo S et al. Systematic identification of human mitochondrial disease genes through integrative genomics. Nat Genet 2006;38(5):576–82.
[17] Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nat Genet 2002;31(3):316–9.
[18] Tim Berners-Lee JH, Ora Lassila. The Semantic Web. Sci Am Mag 2001;284:29–37.
[19] Available from: http://en.wikipedia.org/wiki/Eigenvector_centrality# eigenvector_centrality.
[20] Junker BH, Koschutzki D, Schreiber F. Exploration of biological network centralities with CentiBiN. BMC Bioinformatics 2006;7:219.
[21] Bhuvan B, Sougata M. Utilizing resource importance for ranking Semantic Web query results. Semantic Web Databases 2005:185–98.
[22] Kleinberg JM. Authoritative sources in a hyperlinked environment. J ACM 1999;46(5):604–32.
[23] Page SBaL. The anatomy of a large-scale hypertextual {Web} search engine. Comput Netw ISDN Syst 1998;30(1–7):107–17.
[24] Sougata M, Bhuvan B, Pankaj K. Information retrieval and knowledge discovery utilizing a biomedical patent Semantic Web. IEEE Educ Act Dep 2005:1099–110.
[25] Prioritizer. Available from: http://humgen.med.uu.nl/~lude/prioritizer/.
[26] Liu Y et al. An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. PLoS Comput Biol 2006;2(11):e159.
[27] Song L et al. Novel locus for an inherited cardiomyopathy maps to chromosome 7. Circulation 2006;113(18):2186–92.
[28] Ellinor PT et al. A novel locus for dilated cardiomyopathy, diffuse myocardial fibrosis, and sudden death on chromosome 10q25–26. J Am Coll Cardiol 2006;48(1):106–11.
[29] Grzeskowiak R et al. Expression profiling of human idiopathic dilated cardiomyopathy. Cardiovasc Res 2003;59(2):400–11.
[30] Harris MA et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32(Database issue):D258–61.
[31] Kanehisa M et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006;34(Database issue):D354–7.
[32] Karp PD et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 2005;33(19):6083–9.
[33] Joshi-Tope G et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 2005;33(Database issue):D428–32.
[34] Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 2005;6(1):R7.
[35] Hamosh A et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;33(Database issue):D514–7.
[36] Jablonski S. Jablonski's dictionary of syndromes & eponymic diseases. 2nd ed. Krieger Publication; 1991.
[37] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17–21.
[38] Butte AJ, Kohane IS. Creation and implications of a phenome–genome network. Nat Biotechnol 2006;24(1):55–62.
[39] Cunningham H. GATE, a general architecture for text engineering. Comput Humanit 2002;36(Number 2):223–54.
[40] Louie B et al. Data integration and genomic medicine. J Biomed Inform 2007;40(1):5–16.
[41] Qu XA et al. Semantic Web-based data representation and reasoning applied to disease mechanism and pharmacology. In: Second IEEE international workshop on data mining in bioinformatics (DMB 2007). Silicon Valley, USA; 2007.
[42] Rubin DL, Noy NF, Musen MA. Protege: a tool for managing and using terminology in radiology applications. J Digit Imaging 2007;20(Suppl. 1):34–46.
[43] Noy NF et al. Protege-2000: an open-source ontology-development and knowledge-acquisition environment. AMIA Annu Symp Proc 2003:953.
[44] Mukherjea S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Brief Bioinform 2005;6(3):252–62.
[45] Available from: http://www.w3.org/TR/rdf-sparql-query/#construct.
[46] Ohazama A, Sharpe PT. TFII-I gene family during tooth development: candidate genes for tooth anomalies in Williams syndrome. Dev Dyn 2007;236(10):2884–8.
[47] van Hagen JM et al. Contribution of CYLN2 and GTF2IRD1 to neurological and cognitive symptoms in Williams Syndrome. Neurobiol Dis 2007;26(1):112–24.
[48] Palmer SJ et al. Expression of Gtf2ird1, the Williams syndrome-associated gene, during mouse development. Gene Expr Patterns 2007;7(4):396–404.
[49] Edelmann L et al. An atypical deletion of the Williams–Beuren syndrome interval implicates genes associated with defective visuospatial processing and autism. J Med Genet 2007;44(2):136–43.
[50] Hinsley TA et al. Comparison of TFII-I gene family members deleted in Williams–Beuren syndrome. Protein Sci 2004;13(10):2588–99.
[51] Casanelles Mdel C et al. Portal hypertension in Williams syndrome: report of two patients. Am J Med Genet A 2003;118(4):372–6.
[52] Piontkivska H et al. Multi-species sequence comparison reveals dynamic evolution of the elastin gene that has involved purifying selection and lineage-specific insertions/deletions. BMC Genomics 2004;5(1):31.
[53] Tassabehji M, Urban Z. Congenital heart disease: molecular diagnostics of supravalvular aortic stenosis. Methods Mol Med 2006;126:129–56.
[54] Hoogenraad CC et al. LIMK1 and CLIP-115: linking cytoskeletal defects to Williams syndrome. Bioessays 2004;26(2):141–50.
[55] Morris CA et al. GTF2I hemizygosity implicated in mental retardation in Williams syndrome: genotype–phenotype analysis of five families with deletions in the Williams syndrome region. Am J Med Genet A 2003;123(1):45–59.
[56] Pankau R et al. Familial Williams–Beuren syndrome showing varying clinical expression. Am J Med Genet 2001;98(4):324–9.
[57] Cavellan E et al. The WSTF-SNF2h chromatin remodeling complex interacts with several nuclear proteins in transcription. J Biol Chem 2006;281(24):16264–71.
[58] Kitagawa H et al. The chromatin-remodeling complex WINAC targets a nuclear receptor to promoters and is impaired in Williams syndrome. Cell 2003;113(7):905–17.
[59] Poot RA et al. Chromatin remodeling by WSTF-ISWI at the replication site: opening a window of opportunity for epigenetic inheritance? Cell Cycle 2005;4(4):543–6.
[60] Bruno E et al. Cardiovascular findings, and clinical course, in patients with Williams syndrome. Cardiol Young 2003;13(6):532–6.
[61] Lavine KJ et al. Endocardial and epicardial derived FGF signals regulate myocardial proliferation and differentiation in vivo. Dev Cell 2005;8(1):85–95.
[62] Dzimiri N et al. Differential functional expression of human myocardial G protein receptor kinases in left ventricular cardiac diseases. Eur J Pharmacol 2004;489(3):167–77.

[63] Jahns R et al. Direct evidence for a beta 1-adrenergic receptor-directed autoimmune attack as a cause of idiopathic dilated cardiomyopathy. J Clin Invest 2004;113(10):1419–29.
[64] Stork S et al. Stimulating autoantibodies directed against the cardiac beta1-adrenergic receptor predict increased mortality in idiopathic cardiomyopathy. Am Heart J 2006;152(4):697–704.
[65] Archacki SR, Wang QK. Microarray analysis of cardiovascular diseases. Methods Mol Med 2006;129:1–13.
[66] Jefferies JL et al. Genetic predictors and remodeling of dilated cardiomyopathy in muscular dystrophy. Circulation 2005;112(18):2799–804.
[67] Inagaki N et al. Alpha B-crystallin mutation in dilated cardiomyopathy. Biochem Biophys Res Commun 2006;342(2):379–86.
[68] Kolcz J et al. The expression of connexin 43 in children with Tetralogy of Fallot. Cell Mol Biol Lett 2005;10(2):287–303.
[69] Kannankeril PJ et al. Mice with the R176Q cardiac ryanodine receptor mutation exhibit catecholamine-induced ventricular tachycardia and cardiomyopathy. Proc Natl Acad Sci USA 2006;103(32):12179–84.
[70] Jona I, Nanasi PP. Cardiomyopathies and sudden cardiac death caused by RyR2 mutations: are the channels the beginning and the end? Cardiovasc Res 2006;71(3):416–8.
[71] Clarke AR. Murine genetic models of human disease. Curr Opin Genet Dev 1994;4(3):453–60.
[72] Goh CS et al. Integration of curated databases to identify genotype–phenotype associations. BMC Genomics 2006;7:257.
[73] Cantor MN, Lussier YA. Mining OMIM for insight into complex diseases. Medinfo 2004;11(Pt 2):753–7.
[74] Blake JA, Bult CJ. Beyond the data deluge: data integration and bio-ontologies. J Biomed Inform 2006;39(3):314–20.
[75] Sioutos N et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40(1):30–43.
[76] Osborne JD et al. Mining biomedical data using MetaMap Transfer (MMtx) and the Unified Medical Language System (UMLS). Methods Mol Biol 2007;408:153–69.