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.

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).