Semantic web for integrated network analysis in biomedicine

Huajun Chen, Li Ding, Zhaohui Wu, Tong Yu, Lavanya Dhanapalan and Jake Y. Chen

Submitted: 13th August 2008; Received (in revised form): 6th January 2009

Abstract

The Semantic Web technology enables integration of heterogeneous data on the World Wide Web by making the semantics of data explicit through formal ontologies. In this article, we survey the feasibility and state of the art of utilizing the Semantic Web technology to represent, integrate and analyze the knowledge in various biomedical networks. We introduce a new conceptual framework, semantic graph mining, to enable researchers to integrate graph mining with ontology reasoning in network data analysis. Through four case studies, we demonstrate how semantic graph mining can be applied to the analysis of disease-causal genes, Gene Ontology category cross-talks, drug efficacy analysis and herb–drug interactions analysis.

Keywords: *Semantic Web; network biology; network medicine; graph mining; biomedical network analysis*

NETWORK BIOLOGY AND MEDICINE IN THE SEMANTIC WEB ERA

Recent advances in biomedical research have led to an influx of large amount of association data interlinking genes, proteins, genetic variations, chemical compounds, diseases and drugs [1, 2]. These associations have enabled the emerging studies in Network Biology [3–5], in which the relations between biological molecules and phenotypes are studied from a global perspective. The applications of network biology may lead to future Network Medicine [6], in which multitude of drug targets and panel biomarkers may replace single drug target or single molecular biomarker to improve drug discovery and biomarker development.

A fast developing trend in biomedical network analysis is to combine multiple biomedical association data, which can be highly heterogeneous, into coherent bio-molecular interaction networks to enable integrated network analysis [7–11]. To support challenging biomedical data integration efforts, a computational technique should satisfy the following three basic requirements:

(i) An interoperable data model that is capable of capturing and modeling the observed biomedical networks.
(ii) A data integration framework to map and merge network data across disparate data sources.
(iii) A collection of computational services to analyze, discover and validate new associations in integrated biomedical networks.

Corresponding author. Dr Huajun Chen. Tel: 86-571-87953703; Fax: 86-571-87953079; E-mail: huajunsir@gmail.com

Huajun Chen is an associated professor of School of Computer Science, Zhejiang University, and the deputy director of Center for Traditional Chinese Medicine Informatics, China Academy of Chinese Medicine Sciences (CACMS) and Zhejiang University.

Li Ding is a research scientist of Tetherless World Constellation, Rensselaer Polytechnic Institution. He was previously a postdoctoral fellow of Knowledge Systems, Artificial Intelligence Laboratory (KSL), Stanford University.

Zhaohui Wu is a full professor of School of Computer Science, Zhejiang University, and the director of the Center for Traditional Chinese Medicine Informatics, China Academy of Chinese Medicine Sciences (CACMS) and Zhejiang University.

Tong Yu is a PhD candidate at School of Computer Science, Zhejiang University.

Lavanya Dhanapalan is a PhD Candidate of Department of Computer and Information Science of Purdue University.

Jake Y. Chen is an assistant professor of Informatics and Computer Science, Indiana University School of Informatics, and Purdue University Department of Computer & Information Science. He is the founding director of Indiana Center for Systems Biology and Personalized Medicine, and also an advisory committee member of IU School of Medicine Translational Genomics Core IU Center for Environmental Health.

The Semantic Web [12], proposed by Tim Berners-Lee—the inventor of the World Wide Web, has promising application potentials in life sciences and healthcare [13–17, 79].

First, the Resource Description Framework (RDF Concepts and Syntax: http://www.w3.org/TR/rdf-concepts/), a graph-theoretic data model defined as a web technology standard, is ideally suitable for representing biomedical networks [18–20]. For example, in Figure 1, each biomedical entity (e.g. a gene, a disease, a drug, a person, etc.) is mapped to a node and uniquely identified by a Uniform Resource Identifier (URI: Generic Syntax,http://www.ietf.org/rfc/rfc2396.txt). The connections (e.g. part-of, increase and block) between biomedical entities are captured by certain graph structures, such as a *labeled directed link*. The Semantic Web ontology languages, i.e. RDF

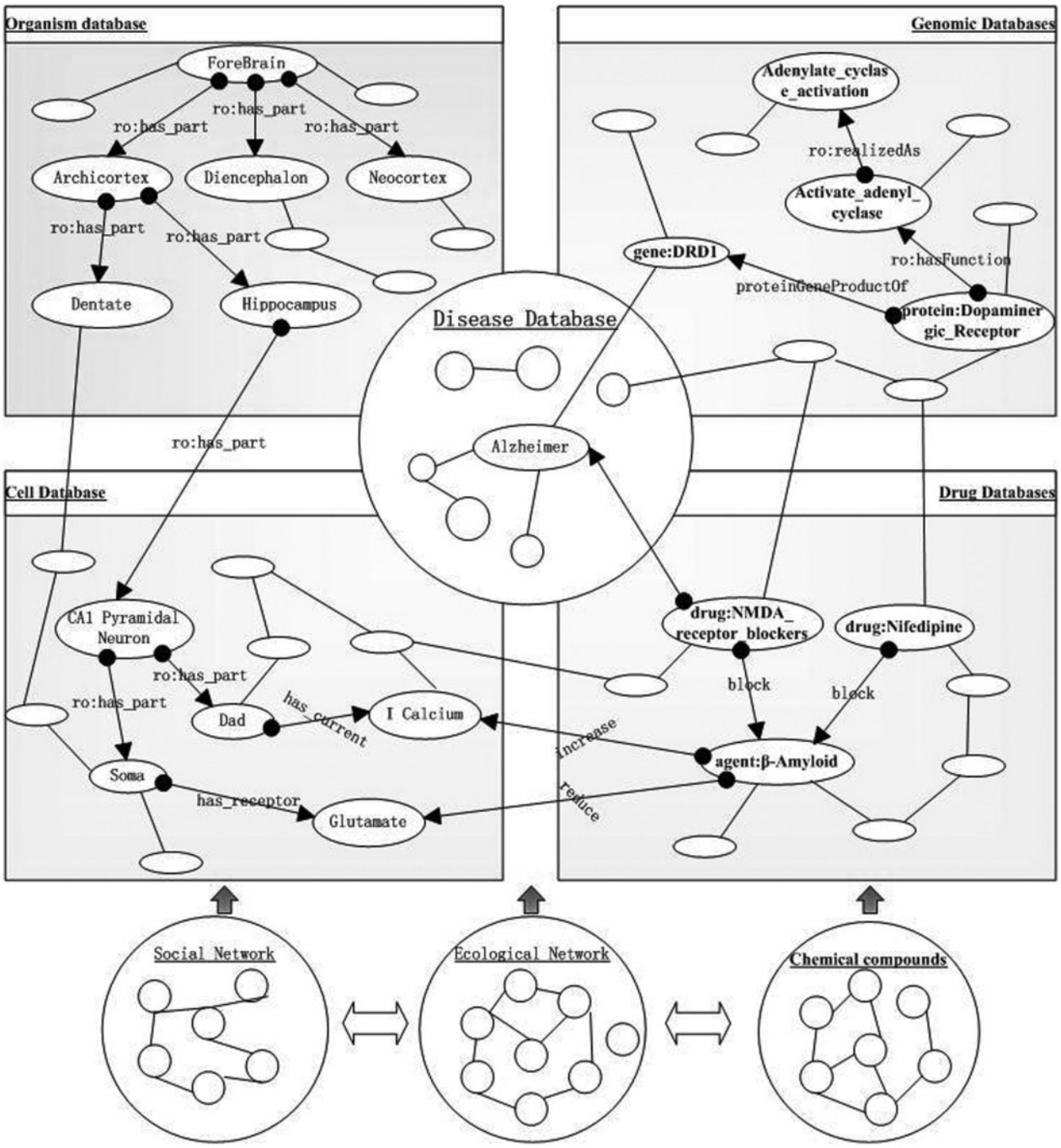

Figure 1: A semantic graph is a web of things, in which both nodes and links are uniquely identified by URIs. A semantic graph can connect data from different sources and domains while preserving the provenance of data.

Schema (RDFS Schema: http://www.w3.org/TR/rdf-schema) and Web Ontology Language (OWL Web Ontology: http://www.w3.org/TR/owl-ref/), further allow one to enrich biomedical networks with formal vocabularies. The result graph is not merely a topological network, but also a 'meaning-preserving network', i.e. whose nodes and links are formally defined in corresponding ontologies.

Secondly, RDF has many innate features that are particularly designed for reducing difficulties in linking and integrating data drawn from disparate sources. For example, the use of open URIs offers formal and standard mechanisms for merging entities and relations across different domains; the RDF triple model (<subject, predicate, object>) is suitable and straightforward for describing relations among entities; and the common and formal vocabularies enabled by Semantic Web ontology languages are the keys to resolving heterogeneity issues across domains.

Thirdly, an RDF-based data graph, in connection with its associated RDF Schema or OWL ontologies, is essentially a knowledge base. We can extend conventional graph mining techniques [22] to utilize the preserved semantics and the reasoning capability of the knowledge base. Typical reasoning computations include consistency checking, subsumption reasoning, entailment inference and rule-based inferences.

Many biomedical networks, especially those collected by the W3C Semantic Web Health Care and Life Sciences Interest Group (HCLS) (http://www.w3.org/2001/sw/hcls), have been represented by Semantic Web languages. For example, the BioRDF task force of HCLS group has hitherto integrated 15 distinct data sources. The result giant biomedical network extensively covers hypothesis, genome, pathway, molecular properties, disease, etc., from molecular level to behavior level, and is growing quickly in active ongoing research.

In the rest of this survey, we show that Semantic Web technologies can be utilized to represent, integrate and analyze the knowledge in various biomedical networks. We show several case studies, which typically encoded biomedical networks using Semantic Web-based graph-theoretic data structures. We summarize the methods and techniques that these studies adopt as a methodology called *Semantic Graph Mining (SGM)*. Different from conventional graph mining approaches [22], SGM deals with Semantic Graphs that encode more semantic signature (such as the labels of nodes and edges) and semantic implication (such as new edges derived from ontologocal inference). SGM is thus characterized by its capability of incorporating graph mining [22] and ontology reasoning [21].

SEMANTIC GRAPH

As many biomedical networks can be typically structured as connected biomedical entities linked by different types of relations, the labeled directed graph model, which is the fundamental data model of the Semantic Web, is indeed more suitable in capturing the knowledge in the networks. In what follows, we briefly review the semantic graph data models, and clarify the definition and properties of the model.

Semantic Web data models and languages

RDF

RDF is the fundamental data model, recommended by W3C, for encoding and sharing data semantics on the Web. RDF provides a simple graph data model for encoding networked data on the Web using node and binary relations. The key ideas of RDF are as follows:

- A web resource (including entity, relation, class, etc.), e.g. a person, a publication, a molecule, a class of genes and a membership relation, is uniquely identified by a URI, which can be used as a global identifier across system or domain boundaries. For example, we can use <http://example.org/flower#Rose› to refer the 'plant' sense of the term 'rose'.
- Information is encoded as RDF triples, i.e. binary relations in the form of ‹*s, p, o*›, which means that a resource *s* holds certain relation *p* with another resource *o*. For example, we can use a triple ‹*ex:Deer, ex:hasPredator, ex:Wolf*› to encode a relation in a food web.
- Each triple can be viewed as an annotated link where the start node, the end node and the link are labeled with unique identifiers (URIs) or text strings. RDF triples can be easily aggregated and combined into a global RDF graph by merging nodes annotated with the same URIs.
- Similar to SQL for relational database, W3C also recommends SPARQL (http://www.w3.org/TR/rdf-sparql-query/), the standard query language for RDF, to access and query data encoded in RDF triples.

RDF Schema and OWL

RDF provides a simple way to model relations between web resources. However, RDF itself provides no means for defining the vocabularies (terms) users intend to use, for example, the vocabularies that describe specific kinds or classes of resources or specific properties in describing those resources. Instead, such classes and properties are described as an RDF vocabulary, using extensions to RDF provided by the RDF Vocabulary Description Language: RDF Schema.

RDF Schema enhances RDF with light weight ontology constructs inherited from semantic networks and frame systems [21]. Similar to frame systems, RDF Schema organizes knowledge via object-oriented modeling and offers built-in semantics such as class hierarchy. RDF Schema also in debts to semantic network by making individual, class, and relation first class citizens; therefore, relations can be independently declared and related via special relations such as rdfs:subPropertyof.

OWL succeeds RDF Schema with well-defined richer set of semantics inherited from description logics [21]. OWL offers set-theoretic ontology constructs such as owl:disjoinWith and owl: intersectionOf, equality relations such as owl: sameAs, owl:equivelantClass, more descriptions describing classes such as cardinality restriction and enumerative class, more special properties associating relations such as owl:TransitiveProperty and owl: FunctionalProperty.

Named Graph

Named Graph [24] extends the RDF graph model by adding notion of context. Each named graph contains a finite RDF graph as its content and a URI as its name. Users can create a named graph to enclose a finite set of RDF triples and refer the named graph using its URI. Data collected from different sources can be enclosed as different named graphs.

The main benefit of the Named Graph is to allow provenance information (e.g. when, where and who created the data) to be associated with data resources. Web agents can use a trust model to calculate the belief of data resources based on data provenance, which grants biomedical researchers finer control over different data sets with the capability of tracking data provenance and reasoning about uncertainty.

Model networks as Semantic Graphs

Based on common practices that model biomedical networks and the graph-theoretic features of RDF data models, we formally define Semantic Graphs as follows:

Definition 1: Semantic Graph. A Semantic Graph g is a set of labeled nodes connected by a set of labeled links. Thus `g = (Name,N,R,L)`, where:

- *Name* is a URI for identifying g, *i.e, the name for the graph*.
- N is a set of labeled nodes. A node *n* is a labeled node and *n::=‹node.label, type.labels›*, where *node.label* is a URI or an RDF literal which is simply natural language string, and the *type.labels* is a set of types annotated on *n, via RDF:type descriptions*. A node may have more than one type annotations.
- R is set of relations, each corresponds to an RDF property in g, i.e. a binary relation R::=(N x N). A relation is labeled by a URI.
- L is a set of labeled links. set of links, each corresponds to an RDF triple (s, p, o), stating the fact that a node s is linked to another node o via a binary relation p.

In Figure 1, four biomedical networks are represented in four corresponding semantic graphs. Each semantic graph has a set of unique nodes labeled by their types corresponds to classes defined in a certain ontology. The link 'has-part' has been reused linking nodes through the four semantic graphs.

Semantic Graphs carry two levels of graph equivalence semantics: (i) on syntactic level, two semantic graphs are equivalent when there exists a well-formed bi-jection between their nodes and their triples can be bi-directionally mapped; (ii) on semantic level, two semantic graphs, which are considered as knowledge base, are equivalent if they semantically entail each other (RDF Semantics http://www.w3.org/TR/rdf-m). In consequence, a sub-graph is no longer a subset of a graph. Instead, an inference engine can perform a *subsumption reasoning or entailment inference* to decide whether one graph, *g1, is logically implied by another one, g2*, to determine whether g1 is a sub graph of g2.

Definition 2: Semantic sub-graph. In a semantic graph G, every transaction can be represented as a knowledge base consisting of statements. One graph ***A*** is a sub-graph of graph ***B*** iff. ***A is logically entailed by B***.

Furthermore, domain knowledge and/or personal preference can be used to assign weights to resources and statements in Semantic Graphs [65]. As Semantic Graphs add labels to nodes and links, resources can be weighted and grouped by types of nodes and links. In section Semantic Graph Ranking, we introduce a case study that assigns distinct weights to different types of links to calculate scores of nodes.

Properties of Semantic Graphs

Semantic Graphs have several distinct properties comparing with conventional graph models. We summarize them as follows:

- *Graph integration*. By labeling nodes and links in Semantic Graphs, users can benefit from the following: (i) semantic integration, i.e. multiple Semantic Graphs can be merged via nodes and links sharing the same label; (ii) semantic weighting, i.e. users can assign weight to nodes and links based on the semantics conveyed by their labels (and/or the label of their selected neighbors).
- *Graph data retrieval*. While conventional graphs are typically stored in database tables, Semantic Graphs may be stored in a knowledge base capable of performing inferences. Moreover, semantic graph data could be decentralized in disparate sources as named graphs for data provenance tracking.
- *Graph structure analysis*. Conventional graphs can be converted to a special form of Semantic Graphs by assigning one label to all nodes and one label to all links. Meanwhile, semantic graph structure analysis may be processed in two steps: (i) use the semantics carried by the labels and structure of a semantic graph to derive an (usually weighted) conventional graph (which may even have different topological structure) and (ii) analyze the derived graph using conventional graph analysis tools.
- *Graph computation*. Conventional graph mining mainly draws on statistics-based computation. As Semantic Graphs carry semantic information that additionally declares more applicable computations. For example, part of semantic graphs may be derived via inference; and two semantic graphs can be equivalent at syntactic level according equivalence semantics of RDF graph (See http://www.w3.org/TR/rdf-concepts/#section-graph-equality) or semantic level according RDF entailment semantics (See http://www.w3.org/TR/rdf-mt/#rdf_entail).

SEMANTIC GRAPH MINING AND BIOMEDICAL NETWORK ANALYSIS

In this section, we survey existing research problems on biomedical network analysis and show how they can be addressed by Semantic Graph Mining (SGM). For each of the research problem, we describe a case study to demonstrate the basic ideas and research applications of SGM.

Semantic Graph ranking

Problems

Ranking is commonly used in sorting search results in genomic information retrieval [25, 26] as well as for detecting essential nodes in biological networks [27] such as for lethality analysis [28] and disease-causal gene identification [29]. General ranking computations can be reduced to *Semantic Graph Ranking* as formulated as below.

Definition 3: Semantic Graph Ranking. Given one semantic graph *SG* or a collection *of integrated semantic graphs SGs*, compute various types of numerical ranking (e.g. centrality, relevance and importance) for globally sorting elements of Semantic Graph(s) at different levels of granularity, such as nodes, relations, sub-graphs, and semantic graphs [30].

Semantic Graph Ranking typically utilizes the following characteristics of Semantic Graphs: (i) semantic graphs are interconnected via shared nodes and relations; (ii) different weights are assigned to nodes and relations based on their labels.

Approaches to ranking nodes in Semantic Graphs have been well introduced in Semantic Web literatures [30]. Common approaches include concept structure analysis (e.g. that looks into certain structural features of concepts, such as structural density [31]), PageRank/HITS-based algorithms (e.g. that analyze links and referrals between ontologies [30, 32]) and link analysis (e.g. that considers and prioritizes different types of links [33]).

Ranking at property/relation level was introduced using link analysis [32, 35] and centrality analysis (e.g. that describes the notions of degree centrality, betweenness centrality and the more complex eigenvector centrality [34]).

Ranking at a sub-graph level has been studied in the context of ranking query results in the Semantic Web [37–40]. Ranking at the semantic graph level (each SG can be viewed as a document) has been studied using both content analysis (e.g. that analyzes the objectivity and subjectivity of textual information [31, 36]) and link analysis (e.g. that considers additional constraints like distances or similarities [30, 33]).

Semantic Graph Ranking has also been applied in various biomedical applications, including retrieving semantically related patents from biomedical patent databases [41], and mining disease-causal Genes from a biological network of multiple relationships [29].

Case study: mining disease-causal genes

The prioritization of candidate genes for specific diseases is crucial to the comprehension of underlying pathophysiological mechanisms. R. C. Gudivada and his colleagues use semantic-based centrality analysis to rank disease-causal genes from a Bio-RDF network [29].

The network is integrated from disparate data sources including Gene Ontology (GO) (http://www.geneontology.org/), gene-pathway annotations (compiled from KEGG [42], BioCarta (http://www.biocarta.com/), BioCyc [43] and Reactome [44]), Mouse Genome Informatics (MGI) website (http://www.informatics.jax.org), Online Mendelian Inheritance in Man (OMIM) [45], and the Multiple Congenital Anomaly/Mental Retardation database [46]. The result integrated network captures genes' relations to the characteristics and symptoms of the condition and similar diseases. A variant ranking algorithm of HITS [47] is then applied on the semantic graph to prioritize hundreds of genes that might be involved with cardiac function.

The proposed algorithm defines two basic semantic metrics to estimate the importance of resources: subjectivity score (SS) and objectivity scores (OS) analogous to hub and authority scores proposed by Kleinberg [47]. Nodes as subject/object of *many* RDF triples have correspondingly *high* subjectivity/objectivity scores. The OS of a node increases if it is pointed to by a node with high SS. Likewise, the SS of a node increases if it points to nodes with high OS.

The study then assigns distinct weights to different types of relations, since the relations are not equally important for disease gene ranking. Specifically, each property is assigned with a *subjectivity weight* (weight that depends on the subject of the property) and an *objectivity weight* (weight that depends on the object of the property). For example, considering the property *inpathway* that relates a gene to a pathway, a gene associated with multiple pathways is more important than a pathway associated with multiple genes. On the other hand, a disease associated with multiple drugs via the property *associateddrug* is less important than a drug associated with multiple diseases. Therefore, properties like *inpathway* have higher subjectivity weight, whereas properties like *associateddrug* have higher objectivity weight. In Figure 2,

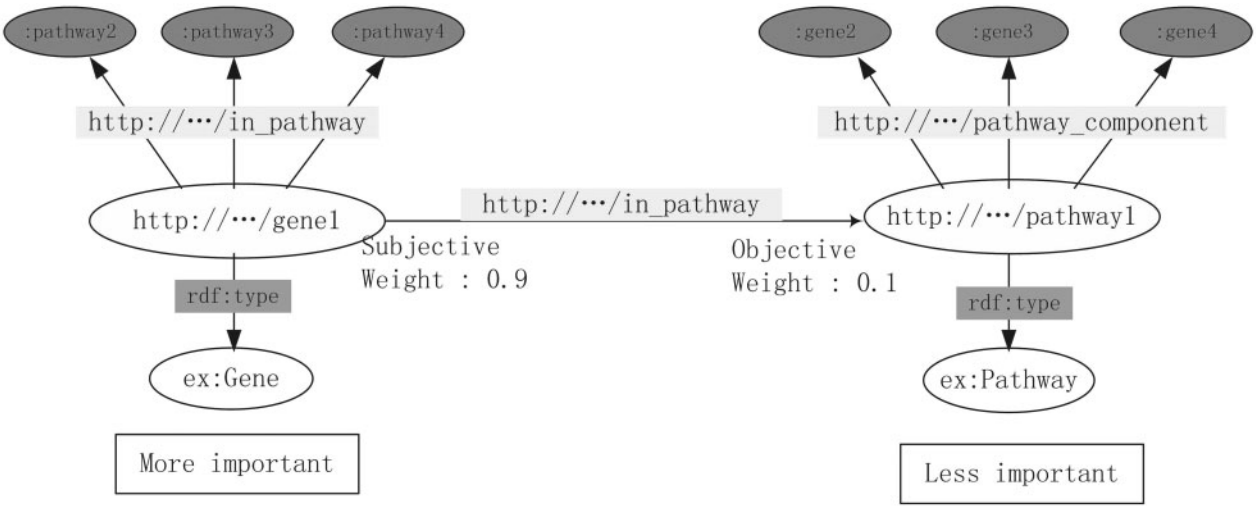

Figure 2: The importance of biomedical entity can be measured by the topology of semantic graph network. A gene associated with multiple pathways (left) is considered more important than a pathway associated with multiple genes (right).

gene is the subject for all the triples and each property is assigned with a subjectivity weight (SW) of 0.9 and objectivity weight (OW) of 0.1. These weights are then used to calculate the centrality of genes. As the result, four candidate genes have been identified with a strong connection to a chromosomal region implicated in dilated cardiomyopathy, a weakening of the heart's pumping ability.

In summary, this case study demonstrates that *SGM* can potentially accelerate the finding of disease-causal genes. First, the Semantic Graphs are integrated from various data sources in a more flexible and efficient manner. Second, as a methodological improvement, context knowledge and user preference can be more easily incorporated into the mining process via semantic weighting to produce more context-sensitive results. More description about these semantic metrics and the experimental result can be found in the original paper [29].

Semantic association discovery

Problems

The purposes of association discovery are *path finding* and *latent association prediction*. Path finding has been used to infer metabolic pathways in metabolite networks [48]. Association prediction is commonly used in protein partnership discovery in analyzing protein interaction networks [49–51] and lethal genetic interactions prediction (e.g that predicts genes working together to control essential functions [52]).

Latent association prediction can be approached by analyzing existing semantic associations between arbitrary resources within the same Semantic Graph or from related Semantic Graphs [37]. One simplest type of semantic association is *transitive association*.

Definition 4: Transitive Association. Two resources r_1 and r_2 are transitively associated if there is a direct path from r_1 to r_2 or from r_2 to r_1.

For example, in Figure 1, the resources *ForeBrain* and *CA1 Pyramidal Neuron* are transitively associated. In the RO ontology (an ontology of core relations for use by OBO Foundry ontologies [53]), *ro:hasPart* is defined as a transitive property. An OWL reasoner can be employed to infer out a further *ro:hasPart* association between *ForeBrain* and *CA1 Pyramidal Neuron* based on the existing *hasPart* relations between *ForeBrain, Archicortex* and *Hippocampus*.

Definition 5: Join Association. Two nodes r_1 and r_2 are joinedly associated (1) if a path from r_1 and another path from r2 can join in the third node, or (2) if two paths from a node can go to both r_1 and r_2.

For example, if two proteins are both participants of a biological process, they are jointly associated by the process. An inference engine may infer a further protein-interaction association between them.

The semantic association discovery problem can be generally framed as follows.

Definition 6: Semantic Association Discovery. Given the integrated semantic graphs SGs and two nodes n_1 and n_2, in terms of association types or strength, how these two nodes are related based on associations existing among others resources.

In what follows, we introduce a case study for GO category cross talk analysis to illustrate the usefulness of semantic information in association discovery [54].

Case study: GO category cross talk analysis

This case study aims at integrating the GO with the network of protein interactions to (1) observe the interactions between proteins utilizing the associations between GO terms, and (2) use protein interactions to infer associations between GO categories at any level [54].

The conventional approach attempts to fit the GO structure, essentially a directed, acyclic graph of GO terms, into a relational database. The performance drawback of this approach is that the GO structure has to be flattened into relational tables for persistent storage, and be rebuilt into a graphical structure upon the retrieval of the parental information for a particular GO term. This structural flattening and rebuilding process is not only time-consuming but also causes data redundancy.

A Semantic Graph enables both straightforward and effective representation of GO and easy integration of GO with protein interaction networks. This study integrates two Semantic Graphs, one for the GO and another for the Human Annotated and Predicted Protein Interactions (HAPPI) database [55]. Figure 3 illustrates a schematic representation of the analysis process including the following steps:

1. Represent GO in RDF format. GO is available in both relational MySQL database and in Semantic Graph (RDF format). The Semantic Graph is chosen for the inherent hierarchical nature of the GO to be retained.

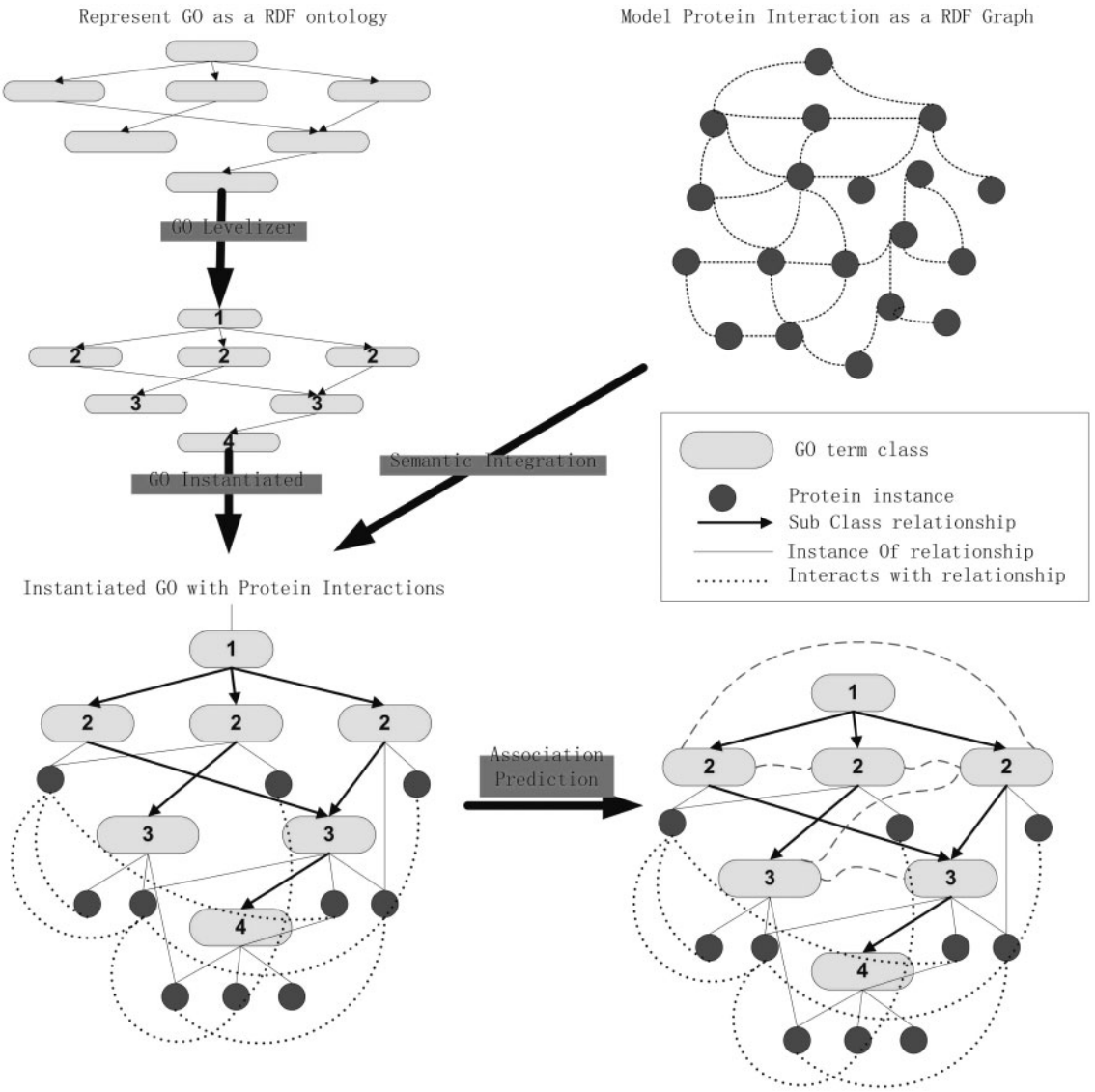

Figure 3: Process of integrating protein interaction dataset with Gene ontology for GO category cross talk analysis.

2. Annotate each GO term with a GO level number, which is defined as the length of the shortest path from the root to the GO term.
3. Model protein network as a Semantic Graph. HAPPI is an Oracle powered relational database. An RDF schema (Table 1) is developed to map and transform the HAPPI data to RDF format.
4. Interlinking the above two Semantic Graphs by linking lower level GO terms to proteins that may map to them. This step is an instantiation of the GO terms with its component proteins that belong to the same GO category. This is the first step towards semantic data integration where the proteins are placed in the context of their GO terms.

Table 1: Mapping the GO cross-talk schema to Semantic graph

Data entities	Link example of Semantic graph
GO terms	(GO:*go.term*, rdf:subClassof, GO:*go.term*)
GO levels	(GO:*go.term*, GO:*hasLevel*, XSD:integer)
Protein instances	(GO:*protein*, rdf:type, GO:*go.term*)
Protein interactions	(GO:*protein1*, GO:*interactsWith*, GO:*protein2*)

5. Use the interactions between proteins in the seed list to calculate the count of the GO-GO category interactions at any specified GO level. For example, upon the request for all associations between Cellular Components at level 3, the

system finds 16096 associations in total at this level, which can be aggregated into 345 distinct GO pairs with association count. The top association is between GO_0044424 (intracellular part) and GO_0043229 (intracellular organelle) that appears 1507 times.

This case study demonstrates that the Semantic Graph can facilitate cross-domain analysis, in which the knowledge discovered in one domain can be reused to analyze the data in another domain. It also demonstrates that native reasoning capability of a semantic graph can be leveraged to boost productivity in mining process. In specific, note that an RDFS reasoner built on the RDF stores (such as Jena) manages all reasoning over sub class, sub property and instance entailments. Therefore, the GO category of a protein at any level can be identified automatically. The efficiency improvement is that it is sufficient to map the proteins to lowest-level GO Terms in the hierarchy without manually tracing back to all the GO ancestral relationships. For details of the results, refer to [54].

Frequent Semantic sub-graphs discovery

Problems

Frequent sub-graphs, commonly referred to as network motifs [56, 57] in biological networks, are a small set of characteristic patterns that occur much more frequently in biomedical networks than in randomized networks. Sub-graphs discovery has been used in discovering regulatory and signaling circuits from gene regulatory networks [58, 59], transcriptional regulatory networks [60] and metabolic pathways [61]. It has also been used in functional analysis for protein interaction networks [62] and gene co-expression networks [63]. Algorithms and tools have been well introduced for the estimation of sub-graph concentrations, and the detection of network motifs [57, 64, 65].

Frequent sub-graphs discovery in Semantic Graphs is interesting because Semantic Graphs carries graph equivalence semantics on both syntactic level and semantic level. As introduced in Definition 2, the semantic level notion of sub-graph requires the incorporation of a reasoning process (subsumption reasoning and entailment inference) in discovering frequent sub-graphs.

The new notion of defining a sub-graph as in Definition 2 brings up another issue as to the definition of the *frequency of a sub-graph* g occurring in a given semantic graph G. We then introduce the concept of 'minimum entailed sub-graph'.

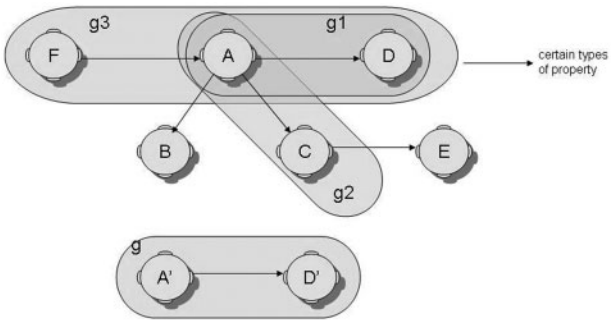

Figure 4: Examples of minimum entailed sub-graph, where each circle represents a resource, each directed arc represents a relation between two resources, and each rectangle with round corn represents a named (sub-)graph. In this example, g is entailed by g1, g2 and g3. g1 and g2, but not g3, are minimum entailed sub-graphs of g. Note that the boxes represent resource identifiers, the lines represent statements, and the ovals represent named sub-graphs.

Definition 7: Minimum Entailed Sub-graph (MES). Give a graph g, and a semantic graph G. A sub-graph g1 of G is a minimum entailed sub-graph with respect to g, iff. (1) g is entailed by g1, (2) there is no other sub-graph g2 of G such that g is entailed by g2 and g2 is entailed by g1.

Intuitively, a MES g1 is a smallest sub-graph of G that entails g. Figure 4 illustrates the basic idea of MES. There might be more than one MES with respect to g. We then introduce the new notion of the frequency of a semantic sub-graph.

Definition 8: Frequent Semantic Sub-Graph. Give a graph g, and a semantic graph G. g is a frequent sub-graph with respect to G, iff. there are more than $\mathbf{i}|K|$ minimum entailed sub-graphs in G with respect to g, where $\mathbf{i}$ is a user-specified minimum support threshold, and $|K|$ is the total number of graphs in K.

Case study: drug efficacy analysis

A typical question in network medicine is to study the associations between drugs and their efficacy. In the scenario illustrated in Figure 5, an analyst retrieves a set of EMRs related to a syndrome named 'Kidney Deficiency (KD)' to analyze the efficacy of KD related drugs. Here, information from each EMR, including a patient's condition (e.g. syndromes and symptoms) and administered drugs and formulae, is represented as a Semantic Graph.

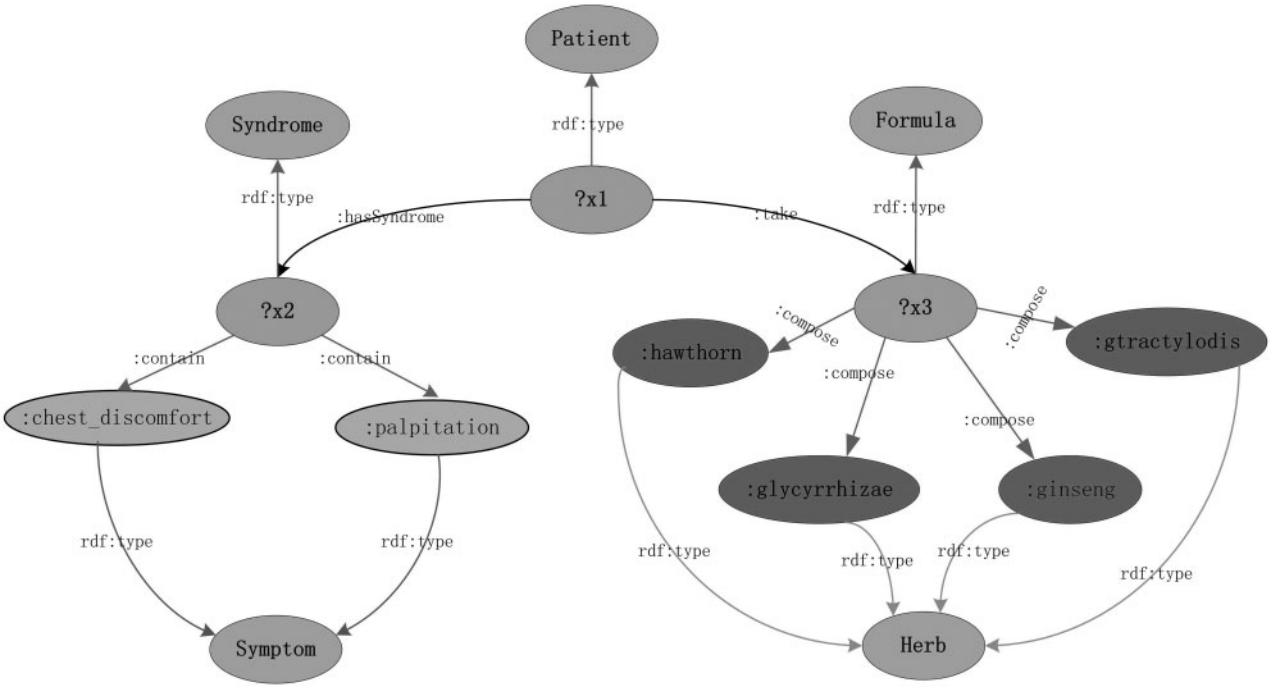

Figure 5: An example of frequent semantic sub-graph illustrating: the formula including the four herbs (right) are frequently used to treat patient having the syndrome with the two symptoms (left). In this example, ?xl, ?x2 and ?x3 are variables.

We develop our approach based on existing frequent pattern mining operators that have been well introduced in data mining literatures [67]. The idea is to treat each link of the Semantic Graph (represented as RDF triple) as an item, then feed those RDF triples into a frequent pattern mining operator to identify frequently occurring sets of triples (sub-graphs). The only difference is that when determining whether a graph is a sub graph of another one, we feed those triples to an OWL reasoner [such as the Racer (http://www.racer-systems.com/index.phtml)] to evaluate the entailment relations.

Figure 5 shows an example frequent semantic sub-graph discovered by this approach. The frequent co-occurrence of four herbs and two symptoms reveals a strong tie that connects the herbs and the symptoms. Here, the discovered pattern (i.e. a sub-graph) is a hypothesis, and each occurrence of the pattern within an EMR counts as an evidence. The hypotheses with their evidences are visualized for expert evaluation. Note that a generalized pattern that contains variables may be associated with matching evidences that instantiate the variable.

This case study demonstrates that Semantic Graph can be used to extract patterns from a great number of medical records, e.g. EMRs. The first advantage is that Semantic Graphs enable the integration of records from different trusted domains. The second advantage is that an OWL-based inference engine can facilitate the mining process with reasoning capabilities such as subsumption reasoning on class hierarchy. For example, suppose we have a class hierarchy for patient that has *FemalePatient* and *MalePatient* as sub-classes of *Patient*. If analysis over female patient data set does not yield satisfying results, SGM supports rolling up to the class *Patient* to include more datasets into the mining input automatically. More complicated RDF entailment inference can be performed to determine whether a sub-graph entails another sub-graph when counting the frequency of sub-graphs.

Semantic Graph clustering

Problems

Cellular functions, such as cell-cycle regulation, are carried out by concerted action of functional modules that are small community structures comprised of many preferentially interacting molecules [68]. Cluster analysis is a common methodology for identifying these functional modules [69, 70]. Clustering can be defined as the grouping of

network nodes based on certain type of association between nodes, e.g. node similarity (measured by the sharing of node properties), and node connectivity (measured by node distance). Clustering has been used in gene expression analysis [71–73], in discovering protein complexes (splicing machinery, transcription factors, etc.) [74–76] and dynamic functional units (signaling cascades, cell-cycle regulation, etc.) [70, 77, 78], and in understanding the organization and functionality of metabolic networks [79].

The Semantic Graph (SG) model has certain characteristics that make it suitable for network-based clustering, including the formal representation of multiple heterogeneous relations, assigning weight to relations by reasoning with domain logics, and formal expression of user preference and problem context. In order to perform network clustering based on Semantic Graph model, we need to define the population under investigation as a Class, and represent the domain knowledge within a Knowledge base. In a specific iteration of clustering, we also need to map a target network as a set of semantic associations according to the specification of the problem-solving context. Therefore, we define the Semantic Graph Clustering problem as follows:

Definition 9 (Semantic Graph Clustering): given a population expressed as class P, a knowledge base KB about P, and a problem-solving context PC, generate a partition of P into a set of communities, based on semantic associations between individuals of P that are extracted from KB and related to PC.

This problem contains three sub-problems: first, find all semantic associations between individuals of P that are extracted from KB and related to PC; second, generate a partition of P into a set of communities, such that there is a higher density of interactions within communities than between them; finally, representation, visualization, and interpretation of the clustering results (partitions and communities).

Herein below, we will illustrate these ideas through a case study in pharmaceutical domain [66].

Case study: analyzing herb–drug interactions

Herb–drug Interaction is an identified open problem in integrative medicine, and particularly affects healthcares delivered in China, where Traditional Chinese Medicine (TCM) is widely used and often in combination with Western medicine. First, TCM pharmacists primarily compound prescriptions as mixtures of multiple herbs, and establish a system of TCM herbal formulae as significant patterns of herb community structure. The essential principle is that a formula should embody a proper herb companionship involving hierarchical social relationships, between a single dominant figure, the king herb and a set of subordinate figures such as minister herbs, assistant herbs, and carrier herbs. Second, the administration of TCM herbs and pharmaceutical drugs leads to interactions between herbs and drugs, which may increase or decrease the pharmacological or toxicological effects of either component. Therefore, herb–drug interaction becomes a critical issue in the efficacy and safety analysis of Chinese herbal medicine.

We describe the rationalization of Semantic Graph Clustering through a TCM use case, which aims at analyzing herb–drug interactions in the therapies of cardiovascular system diseases (CVD). As Figure 6 illustrates, in order to map a global network of herb–drug interactions, all CVD-associated herbs/drugs are first discovered, and then all interactions between them are collected. The nature and frequency of interactions between all pairs of drugs are discovered through semantic association discovery using Semantic Graphs collected from EMR prescription, TCM syndrome, TCM disease, orthodoxy medicine disease, functional chemical component, TCM herb classification and herb therapeutic features. The result semantic graph is generated by inserting a statement for every interaction of herbs. The statement represents a pair of herbs as subject and object, the type of interaction as predicate, and frequency as weight. Based on the result Semantic Graph, communities are identified by first deleting a certain number of centrality nodes, and then deleting a certain number of centrality edges, until communities are partitioned. The process of discovering a community and the interactions within the community serve as semantic annotations to interpret the community. The details of this case study are mentioned in [66].

DISCUSSION

In this section, we summarize the benefits and challenges of using Semantic Web technologies in biomedical network analysis.

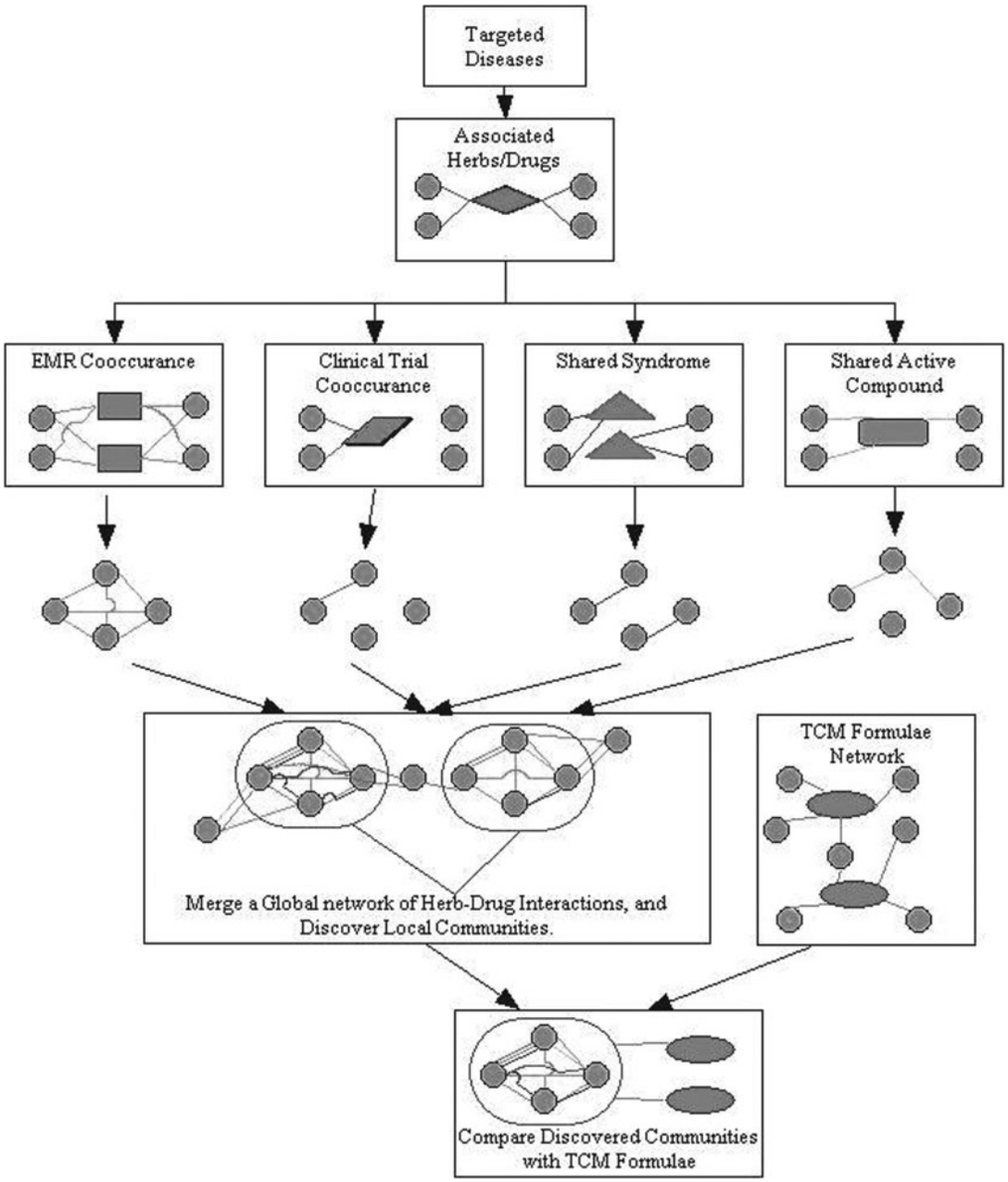

Figure 6: A global network modeling approach for analyzing herb–drug interactions and interpreting TCM formulae system.

Benefits

1. The Semantic Web has the expressive power to capture the semantics of biomedical complex networks. RDF is a simple model to make statements about relations. RDFS and OWL offer more advanced ontology constructs to define classes and properties that can be used to annotate biomedical networks with rich semantics. A Semantic Graph is thus more informative. As the complexity in modeling biological networks increases, the graph model needs to be more expressive in capturing and exhibiting this complexity.
2. The Semantic Web enables integrative analysis of highly heterogeneous networks. Integrating multiple graphs that may be highly heterogeneous is obviously necessary for advanced network biology and network medicine research. RDF is designed and proposed for data integration at a web-scale. The design considerations for using URI, for common

vocabularies and adopting simple graph model make RDF particularly versatile in mapping and integrating heterogeneous networks.
3. The Semantic Web enables ontology reasoning. An RDF graph is essentially a knowledge base. Typical reasoning computations such as consistency checking, subsumption reasoning, entailment inference and rule-based inferences can be performed. As exhibited by several case studies in this article, ontology reasoning can be combined with conventional graph mining approach to enhance mining procedures and improve mining results.
4. The Semantic Web enables tracking of data provenance. Data provenance tracking refers to annotating data with its sources, curation processes, etc. Since many data mining efforts depend on data derived from disparate sources and domains, researchers often need provenance data to evaluate the credibility of discovered results. Named graph enables assigning confidence measures in a model of trust, archiving and versioning RDF data from different sources, summarizing and explaining data provenance in the mining process.

Challenges and limitations

While SGM provides the community with many new features and approaches to advancing biomedical network analysis, it also has the following challenges and limitation.

1. Decentralized mining infrastructure. Semantic Web is a decentralized technology that spans over multiple domains. It brings up additional complexities in decentralized mining, automated data process flow, heterogeneous graph fusion, and data provenance management.
2. Ensuring semantic consistency. Although significant community efforts have been invested in building common semantic model and formal vocabularies for biomedical domain, it is difficult and costly to maintain and ensure consistency between ontologies produced by different communities. This limitation demands more advanced tools and automatic approaches to ease the maintenance of semantic consistency.
3. Graph transformation. While conventional graph mining approaches match well with simple graph networks, each of the integrated Semantic Graphs typically uses its unique types of nodes and relations. Systematic approaches must be developed to transform complex semantic graph(s) into a simple graph network structure to reuse currently available graph manipulation tools. How these transformations affect the interpretation needs thorough understandings.
4. Performance and scalability. As SGM typically involves an ontology reasoning process and needs to deal with huge amount of network data that integrate from multiple sources, it requires high performance computational solutions. As the complexity and scale of biomedical networks under investigation increase over time, the scalability of algorithms will become increasingly critical.

SUMMARY

We surveyed recent biological network analysis and data mining studies based upon the Semantic Web technologies. This technology's strength is demonstrated with case studies to rank disease-causal genes, discover semantic associations based on GO, discover frequent semantic sub-graphs for drug efficacy analysis, and perform semantic graph clustering for drug interaction analysis.

Semantic Graph Mining exhibits several unique characteristics that can support integrated inter-network analysis, provenance-aware decentralized mining, semantic content and link analysis, syntheses of graph mining and ontology reasoning. As the technology evolves over time, remaining challenges regarding decentralized mining infrastructure, ensuring semantic consistency, graph transformation, and performance and scalability require ongoing research and applications in the biological domain.

Key Points

- The Semantic Web offers a knowledge representation and data integration framework that may benefit integrative network biology and network medicine research.
- RDF, RDFS and OWL not only capture the topology of complex biomedical networks, but also preserve the semantics of these networks by annotating nodes and edges with the vocabulary defined in corresponding Semantic Web ontologies.
- *Semantic Graph Mining* is characterized by its capability of integrating graph mining and ontology reasoning for better analyzing biomedical networks.
- Typical SGM computational services include semantic graph ranking, semantic association discovery, frequent semantic sub-graph discovery and semantic graph clustering.
- Named graph enables assigning confidence measures, archiving and versioning data from different sources, summarizing and explaining data provenance in the mining process.

Acknowledgements
We would like to thank all of the colleagues who contributed their ideas to this article.

FUNDING

The researches of the co-authors are funded by China national programs with No. NSFC60525202/NSFC60533040, 2003CB317006, 51306030101, 2006AA01A122 and NSFC60503018; NSF contract 0524481; and DARPA contracts #55-002001, FA8650-06-C-7605.

References

1. Muhammed AY, Goh K-I, Cusick ME, Barabasi A-L, Vidal M. Drug-target network. *Nat Biotechnol* 2007;**25**:1119–26.
2. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A-L. The human disease network. *Proc Natl Acad Sci* 2007;**104**:8685–90.
3. Barabsi AL, Oltvai Z. Network biology: understanding the cells functional organization. *Nat Rev Genet* 2004;**5**:101–13.
4. Albert R. Scale-free networks in cell biology. *J Cell Sci* 2005;**118**:4947–57.
5. Aittokallio T, Schwikowski B. Graph-based methods for analyzing networks in cell biology. *Brief Bioinform* 2006;**7**:243–55.
6. Barabasi A-L. Network medicine-from obesity to the diseasome. *New Engl J Med* 2007;**357**:404–7.
7. Troyanskaya OG. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinform* 2005;**6**:34–43.
8. Tanay A, Sharan R, Kupiec M, *et al*. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome wide data. *Proc Natl Acad Sci USA* 2004;**101**:2981–6.
9. Yeger-Lotem E, Sattath S, Kashtan N, *et al*. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci USA* 2004;**101**:5934–9.
10. Hwang D, Rust AG, Ramsey S, *et al*. A data integration methodology for systems biology. *Proc Natl Acad Sci USA* 2005;**102**:17296–301.
11. Balasubramanian R, LaFramboise T, Scholtens D, *et al*. A graph-theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics* 2004;**20**:3353–62.
12. Berners-Lee T, Handler J, Lassila O. The semantic web. *Scientific American* 2001.
13. Neumann EK. A life science semantic web: are we there yet? *Science, STKE* 2005;**283**:22–5.
14. Neumann EK, Miller E, Wilbanks J. What the semantic web could do for the life sciences. *Drug Disc Today: BIOSILICO* 2006;**2**:228–34.
15. Stephens S, Morales A, Quinlan M. Applying semantic web technologies to drug safety determination. *IEEE Intell Syst* 2006;**21**:82–6.
16. Good BM, Wilkinson MD. The life sciences semantic web is full of creeps! *Brief Bioinform* 2006;**7**:275–86.
17. Ruttenberg A, Clark T, Bug W, *et al*. Advancing translational research with the semantic web. *BMC Bioinform* 2007;**8**(Suppl 3):S2.
18. Wang X, Gorlitsky R, Almeida JS. From XML to RDF: how semantic web technologies will change the design of omic standards. *Nat Biotechnol* 2005;**23**:1099–103.
19. Feigenbaum L, Herman I, Hongsermeier T, *et al*. The semantic web in action. *Scientific American* 2007.
20. Golbreic C. Obo and owl: leveraging semantic web technologies for the life sciences. In: *Proceedings of the International Semantic Web Conference 2007*.
21. Horrocks I, Patel-Schneider PF, van Harmelen F. From SHIQ and RDF to OWL: the making of a web ontology language. *J Web Semant* 2003;**1**:7–26.
22. Chakrabarti D, Faloutsos C. Graph mining: laws, generators, and algorithms. *ACM Comput Surveys* 2006;**38**:Article 2.
23. Stumme G, Hotho A, Berendt B. Semantic web mining state of the art and future directions. *J Web Semant: Sci, Serv Agents World Wide Web* 2006;**4**:124–43.
24. Carroll JJ, Bizer C, Hayes P, *et al*. Named graphs, provenance and trust. In: *Proceedings of the World Wide Web Conference*, 2005.
25. Weston J, Kuang R, Leslie C, *et al*. Protein ranking by semi-supervised network propagation. *BMC Bioinform* 2006;**7**(Suppl 1):S10.
26. Kuang R, Weston J, Noble WS, *et al*. Motif-based protein ranking by network propagation. *Bioinformatics* 2005;**21**:3711–8.
27. Koschtzki D, Schwbermeyera H, Schreibera F. Ranking of network elements based on functional substructures 2007;248:471–9.
28. Palumbo MC, Colosimo A, Giuliani A, *et al*. Functional essentiality from topology features in metabolic networks: A case study in yeast. *FEBS Lett* 2005;**579**:4642.
29. Gudivada RC, Qu XA, Jegga AG, *et al*. A genome-phenome integrated approach for mining disease-causal genes using semantic web. In: *Proceedings of the WWW2007 Workshop on Health Care and Life Science Data Integration for the Semantic Web*, 2007.
30. Ding L, Pan R, Finin T, *et al*. Finding and ranking knowledge on the semantic web. In: *Proceedings of the International Semantic Web Conference*, 2005, pp. 156–170.
31. Alani H, Brewster C. Ontology ranking based on the analysis of concept structures. In: *Proceedings of the 3rd International Conference on Knowledge Capture (K-Cap)*, 2005.
32. Hogan A, Harth A, Decker S. ReConRank: a scalable ranking method for semantic web data with context. In: *Proceedings of the 2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2005.
33. Patel C, Supekar K, Lee Y, *et al*. OntoKhoj: a semantic web portal for ontology searching, ranking and classification. In: *Proceedings of the Conference on Web Information and Data Management*, 2003, pp. 58–61.
34. Hoser B, Hotho A, Jaschke R, *et al*. Semantic network analysis of ontologies. In: *Proceedings of the European Semantic Web Conference*, 2005.

35. Rocha C, Schwabe D, Aragao MP. A hybrid approach for searching in the semantic web. In: *Proceedings of the World Wide Web Conference*, 2004, pp. 374–83.
36. Supekar K, Patel C, Lee Y. Characterizing quality of knowledge on semantic web. In: *Proceedings of the 7th International Florida Artificial Intelligence Research Society Conference*, 2002.
37. Anyanwu K, Maduko A, Sheth A. Semrank: Ranking complex relationship search results on the semantic web. In: *Proceedings of the World Wide Web Conference*, 2005, pp. 117–27.
38. Stojanovic N, Studer R, Stojanovic L. An approach for the ranking of query results in the semantic web. In: *Proceedings of the International Semantic Web Conference*, 2003.
39. Aleman-Meza B, Halaschek C, Arpinar IB, *et al*. Context-aware semantic association ranking. In: *Proceedings of the International Workshop on Semantic Web and Database*, 2003, pp. 33–50.
40. Bhuvan B, Sougata M. Utilizing resource importance for ranking semantic web query results. In: *Proceedings of the International Workshop on Semantic Web and Database*, 2005.
41. Mukherjea S. Information retrieval and knowledge discovery utilising a biomedical semantic web. *Brief Bioinform* 2005;**6**:252–62.
42. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**:354–7.
43. Karp PD, Ouzounis CA, Moore-Kochlacs C, *et al*. Expansion of the Bio-Cyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005;**33**:6083–9.
44. Joshi-Tope G, Gillespie M, Vastrik I, *et al*. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005;**33**:428–32.
45. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**:514–7.
46. Jablonski S. Jablonski's Dictionary of Syndromes & Eponymic Diseases. 2nd edn. Krieger Publishers, 1991.
47. Kleinberg JM. Authoritative sources in a hyperlinked environment. *J ACM* 1999;**46**:604–32.
48. Croes D, Couche F, Wodak SJ, *et al*. Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res* 2005;**33**:326–30.
49. Yu H, Paccanaro A, Trifonov V, *et al*. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* 2006;**22**:823–9.
50. Albert I, Albert R. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics* 2004;**20**:3346–52.
51. Ben-Hur A, Noble WS. Kernel methods for predicting protein-protein interactions. *Bioinformatics* 2005;**21** (Suppl. 1):38–46.
52. Wong SL, Zhang LV, Tong AH, *et al*. Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci* 2004;**101**:15682–7.
53. Smith B, Ceusters W, Klagges B, *et al*. Rosse C relations in biomedical ontologies. *Genome Biol* 2005;**6**:R46.
54. Chen JY, Zhong Y, Changyu S, *et al*. A systems biology approach to the study of cisplatin drug resistance in ovarian cancers. *J Bioinform Comput Biol* 2007;**5**:383–405.
55. Chen JY, Mamidipalli SR, Huan T. HAPPI: an Online Database of Comprehensive Human Annotated and Predicted Protein Interactions. *BMC Genomics* 2009; in press.
56. Milo R, Shen-Orr S, Itzkovitz S, *et al*. Network motifs: simple building blocks of complex networks. *Science* 2002;**298**:824–7.
57. Ciriello G, Guerra C. A review on models and algorithms for motif discovery in protein-protein interaction networks. *Brief Funct Genomics Proteomics* 2008;**7**:147–56.
58. Berg J, Lassig M. Local graph alignment and motif search in biological networks. *Proc Natl Acad Sci* 2004;**101**:14689–94.
59. Ideker T, Ozier O, Schwikowski B, *et al*. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002;**18**(Suppl. 1):S233–40.
60. Koyuturk M, Grama A, Szpankowski W. An efficient algorithm for detecting frequent sub-graphs in biological networks. *Bioinformatics* 2004;**20**(Suppl. 1):I200–7.
61. Jiang R, Tu Z, Chen T, Sun F. Network motif identification in stochastic networks. *Proc Natl Acad Sci* 2006;**103**:9404–9.
62. Hu H, Yan X, Huang Y, *et al*. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics* 2005;**21**(Suppl 1):i213–21.
63. Kashtan N, Itzkovitz S, Milo R. Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs. *Bioinformatics* 2004;**20**:1746–58.
64. Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection. *Bioinformatics* 2006;**22**:1152–3.
65. Schreiber F, Schwobbermeyer H. MAVisto: a tool for the exploration of network motifs. *Bioinformatics* 2005;**21**:3572–4.
66. Tong Y, Chen H. Semantic graph mining for biomedical network analysis. In: *Proceedings of the WWW Workshop on Semantic Web for Health Care and life Science*, 2008.
67. Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. *Data Mining Knowl Discov* 2007;**15**:55–86.
68. Hartwell LH, Hopfiled JJ, Leibler S, *et al*. From molecular to modular cell biology. *Nature* 1999;**402**(Suppl. 6761):C47–52.
69. Newman MEJ. From the cover: modularity and community structure in networks. *Proc Natl Acad Sci* 2006;**103**:8577–82.
70. Rives AW, Galitski T. Modular organization of cellular networks. *Proc Natl Acad Sci* 2003;**100**:1128–33.
71. Dhaeseleer P. How does gene expression clustering work? *Nat Biotechnol* 2005;**23**:1499–501.
72. Segal E, Shapira M, Regev A, *et al*. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003;**34**:166–76.
73. Bar-Joseph Z, Gerber GK, Lee TI, *et al*. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 2003;**21**:1337–42.
74. King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics* 2004;**20**:3013–20.

75. Dunn R, Dudbridge F, Sanderson CM. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMCBioinform* 2005;**6**: 39–47.

76. Farutin V, Robison K, Lightcap E, *et al*. Edge-count probabilities for the identification of local protein communities and their organization. *Proteins* 2006;**62**: 800–18.

77. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci* 2003;**100**: 12123–8.

78. Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. *Proteins* 2004;**54**:49–57.

79. Ma HW, Zhao XM, Yuan YJ, *et al*. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics* 2004;**20**:1870–6.

80. Stephens S, LaVigna D, DiLascio M, Luciano J. Aggregation of bioinformatics data using semantic web technology. *J Web Semant: Sci Serv Agents World Wide Web*;**4**:216–221.