Semantic Rules on Drug Discovery Data

Sashikiran Challa
Indiana University
901 E 10th
Bloomington, IN
+1-812-856-1848
schalla@indiana.edu

David Wild
Indiana University
901 E 10th
Bloomington, IN
+1-812-856-1848
djwild@indiana.edu

Ying Ding
Indiana University
1320 E 10th
Bloomington, IN
+1-812-855-5388
dingying@indiana.edu

Qian Zhu
Indiana University
901 E 10th
Bloomington, IN
+1-812-856-1848
qianzhu@indiana.edu

ABSTRACT

There are several publicly available repositories of chemical compounds and their biological information, like PubChem Compound, PubChem BioAssay, Drug Bank. Presenting all the information about a particular compound and also the related information about all the compounds similar to that particular compound helps scientists make a great progress in the drug discovery. In this paper we present a simple semantic rule based prototype that allows the integration of data procured from different sources according to the Wendi Ontology (an Ontology created in house for this purpose) and make valid and useful inferences that help in drug discovery process.

Categories and Subject Descriptors
I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks

General Terms
Measurement, Design,

Keywords
semantic web, Ontology, RDF data model, generic reasoner

1. INTRODUCTION

In the field of drug discovery there are several diverse repositories of information. Some are publicly available like PubChem Compound, Pubchem BioAssay, Drug Bank or some are proprietary databases in house in the Pharmaceutical industry. There are several Web-Services to specifically fetch either predictive models information or journal articles information. Aggregating information from several different web services is being achieved. If there were a way to not only aggregate information but also make some inferences and establish relationships among the data aggregated, then it would be really helpful to medicinal chemists to explore the full picture of potential biological activity of the compound. The work presented in this paper is towards achieving this using semantic web technology. By having the data in the form of RDF data model and by framing rules that can establish new relationships between the data, this can be achieved. The use of rules in Chemical and Life Sciences areas covers areas such as prediction of oral bioavailability of Chemical Compounds [1], the prediction of human drug metabolism and toxicity of novel compounds [2].

Semantic Web Technologies are being applied in every domain of knowledge and are helping towards building the web of data. The work presented here is not towards the web aspect of the semantic web but towards the Semantics aspect of the data. Data here is the drug discovery data that was collected from WENDI (Web Engine for Non-Obvious Drug Information) web service. WENDI is a tool for finding non-obvious associations between compounds and biology. It is a product of Collaboration between Cheminformatics group at Indiana University, School of Informatics and Computing and Eli Lilly. It takes a single compound as a query and then aggregates comprehensive information about a compound from web services that represent several diverse sources (including predictive models, chemical compound databases, and journal articles). The aggregate information is obtained in the form of an XML. For the work presented only the information about active similar compounds and Journal articles was extracted using XML mini DOM parser in Python scripting language. The extracted information was converted into RDF triples according to an Ontology called WendiOntology that was created in house for this purpose.

WendiOntology: An Ontology named ‘WendiOntology’ was created using Protege4.0. Chemical Compound, BioAssay, Journal Article were the 3 classes defined. Is_similarTo (domain: Chemical Compound, range: Chemical Compound) is_AssociatedWith (domain: Chemical Compound, range: BioAssay), Is_ContainedIn (domain: Chemical Compound, range: Journal Article) were defined as the Object Properties. ‘has_PubMedID’, ‘has_title’ were defined as the Data type properties. Part of the triples (showing single CID and AID relationship) **WendiOntology:cid15940175 rdf:type owl:Thing, owl:ChemicalCompound; WendiOntology:is_AssociatedWith WendiOntology:aid1004**.

2. Framing the Rules

Thus the RDF triples that were generated based on WendiOntology were loaded into Ont Model Class in Jena, a java framework for building semantic web applications. Then rules were written as following in the syntax of SPARQL statements.

"[rule1:(?x http://www.semanticweb.org/ontologies/2009/6/WendiOntology.owl#is_SimilarTo ?y) "+" (?y http://www.semanticweb.org/ontologies/2009/6/WendiOntology.owl#is_ContainedIn ?z) "+" ->(?x http://www.semanticweb.org/ontologies/2009/6/WendiOntology.owl#mightcontaininfo ?z)]";

"[rule2:(?x http://www.semanticweb.org/ontologies/2009/6/WendiOntology.owl#is_SimilarTo ?a) "+ "(?a http://www.semanticweb.org/ontologies/2009/6/WendiOntology.owl#is_AssociatedWith ?b) "+" ->(?x http://www.semanticweb.org/ontologies/2009/6/WendiOntology.owl#mightbeAssociated ?b)]";

Generic Rule reasoner belonging to Reasoner class in Jena was called and these rules were parsed by it to generate RDF triples that were not originally containing properties 'mightcontaininfo', 'mightbeAssociated' defined above. On these additionally generated triples SPARQL queries were written to output the information about the query compound, Assay ids, Journal article titles, and their Pubmed ids.

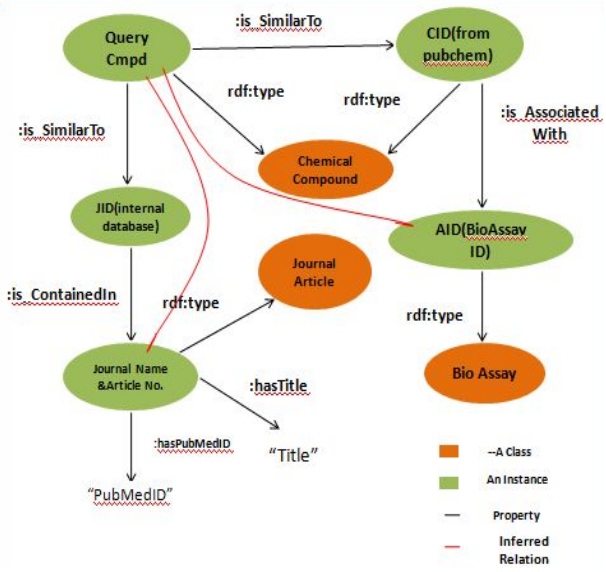

Figure 1: RDF Graph generated on the Drug Discovery Data

3. Conclusion

These are the results that were achieved. Say if a compound A was to found to be similar to compound B and if compound B was found to be active in a Bio Assay C, then a statement that Compound A might be active in Bio Assay C, was a simple inference achieved here. Also say if a compound A was found to be similar to compound B and if compound B was found to be contained in a Journal C, then a statement that compound A's relevant information can be contained in Journal C was another simple inference shown here. Thus the work described here helps us to realize the potential of having the data as RDF data model triples based on an Ontology. It helps a great deal in integrating the data and making useful inferences and thus enhances medicinal chemist's research.

Further work is being done to extend these rules to extract relationships between a disease and its GO terms and genes, which really help medicinal chemists to choose a particular molecule as a drug candidate. Even the inactive or negative relationships are being established for example to show a particular compound is inactive in a particular bioassay and implications it can further make. The edges generated in the RDF graphs are being weighed by giving a probability value to it to let the medicinal chemist know the probability of that relationship that is inferred.

4. References

[1] Stephens, Susie. (2005). "Enabling Semantic Web Inferencing with Oracle Technology: Applications in Life Sciences". *Lecture Notes in Computer Sciences.* 3791, 8-16.

[2] Stepehens, Susie. Morales, Alfredo. Quinlan, Matthew. (2006). "Applying Semantic Web Technologies to Drug Safety Determination". *IEEE Intelligent Systems*, vol.21, no.1, pp.82-86, Jan./Feb.2006, doi:10.1109/MIS.2006.2

[3] Goble, Carole. Stevens, Robert. Bechofer, Sean. (Autumn 2005). "The Semantic Web and Knowledge Grids". *Drug Discovery Today: Technologies*, Vol.2, issue 3, 225-233.

[4]Wang, Xiaoshu. Gorlitsky Robert. Almeida, S, Jonas. (2005). "From XML to RDF: how semantic web technologies will change the design of 'omic' standards." *Nature Biotechnology*, Vol.23, 9, 1099-1103.

[5] Hugo.Y.K.Lam. Marenco, Luis. Clark, Tim. Gao, Yong. Kinoshita, June. Shepherd, Gordon. Miller, Perry. Wu, Elizabeth. Wong, T. Gwendolyn. Liu, Nian. Crasto. Chiquito. Morse, Thomas. Stepehen, Susie. Cheung, Keihoi. (2007) "AlzPharm: integration of neurodegeneration data using RDF." *BMC Bioinformatics*, 8. http://www.biomedcentral.com/1471-2105/8/S3/S4.