Chem2Bio2RDF Dashboard: Ranking Semantic Associations in Systems Chemical Biology Space

Xiao Dong[1] Ying Ding[2] Huijun Wang[1]
Bin Chen[1] David J Wild[1]
[1]School of Informatics and Computing, Indiana University, Bloomington, IN, USA
[2]School of Library and Information Science, Indiana University, Bloomington, IN, USA
{xdong|dingying|huiwang|binchen|djwild}@indiana.edu

ABSTRACT
Semantic Web technology has had a significant impact in scientific collaboration as it provides a common platform to integrate heterogeneous data sources and reasoning capabilities for knowledge discovery. In the biomedical science domain, more and more data providers are providing data in formats that are readily converted to Semantic Web formats, and this has resulted in some early initiatives to collate data in unified Semantic Web repositories such as Linked Open Drug Data (LODD) and Bio2RDF. Many critical problems in biomedical science can be phrased in terms of finding the necessary associations between individual entities (such as explaining drug mechanisms through associations between drugs and metabolic pathways). The networks are necessarily very large, and many association paths may exist between two given entities; therefore an effective and scalable framework for semantic association ranking is needed. In this paper, we describe Chem2Bio2RDF Dashboard, a prototype system for automatic collecting semantic associations within the systems chemical biology space and apply a series of ranking metrics to select the most relevant associations.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Selection process

General Terms
Design, Algorithm

Keywords
Chem2Bio2RDF, Semantic Association Ranking, RDF, SPARQL, Chemical System Biology

1. INTRODUCTION
There is a pressing need in the biomedical science community for effective frameworks to support cross domain data mining. Recent technological and experimental advances have resulted in an explosion of public data about chemical compounds, genomes, biological molecules, and the associated phenotypic and physiological studies. As a direct result, new disciplines have emerged in these interdisciplinary borders including *chemogenomics* and *systems chemical biology* [1]. Chemogenomics studies the impact of chemicals on biological systems in particular interaction among chemical entities and protein molecules, and *systems chemical biology* refers to the integration between cheminformatics and bioinformatics in the realm of systems biology.

The Semantic Web technology offers several technologies that can aid in this process and is thus being adopted in various areas of life sciences, healthcare and drug discovery [2, 3, 4]. Resource Description Framework (RDF) allows data objects, resources and relationships to be described in triple format; Web Ontology Language (OWL) enables conceptualization and classification for data objects and relationships, and it also allows various logical constructs to be specified formally; SPARQL is the query language for RDF data model. Jointly, these semantic web technologies offer rich semantic description, formal annotation and flexible exploration capacities to bring the distributed and heterogeneous biomedical data sources onto the Semantic Web.

2. CHEM2BIO2RDF
Chemogenomics and systems chemical biology, as previously mentioned, are important as they are located at the interface area between biology, medicinal chemistry and drug discovery, and have significant impact over pharmaceutical R&D process. Research studies in these two areas typically involve quite significant collaborative efforts from a number of specialized domains, and hence make it an ideal scenario to deploy Semantic technology. Although earlier initiatives have extended certain coverage over this domain (such as PubChem and DrugBank in LODD[5] and Bio2RDF[6]), there have been no significant efforts to bring integrated chemical and biological data into the Semantic Web domain. We have generated such a resource, called Chem2Bio2RDF [1]. To create this, we collected all the major chemogenomics data sources and combined them with existing biological, phenotypic and systems biology data sources into a single source that currently contains 80 million triples.

In Figure 1 is the linkage scenario for Chem2Bio2RDF. The data sources are arranged into vertical categories of phenotype, chemical, chemogenomics, and systems biology from left to right. The shapes and sizes of polygons refer to the category and proportion of data coverage respectively. Two data sources are connected by a gray edge if common compound is shared and a red edge if common protein (or gene). The diamonds arranged

Copyright is held by the author/owner(s).
FWCS2010, April 26, 2010, Raleigh, USA.

[1] http://chem2bio2rdf.org/ [2] http://www.kegg.org/
[3] http://ctd.mdibl.org/ [4] http://www.bindingdb.org/bind/index.jsp
[5] http://www.pharmgkb.org/ [6] http://matador.embl.de/
[7] http://www.cheminformatics.org [8] http://www.drugbank.ca/
[9] http://www.ncbi.nlm.nih.gov/pcassay
[10] http:www.ncbi.nlm.nih.gov/pubmed/11752352

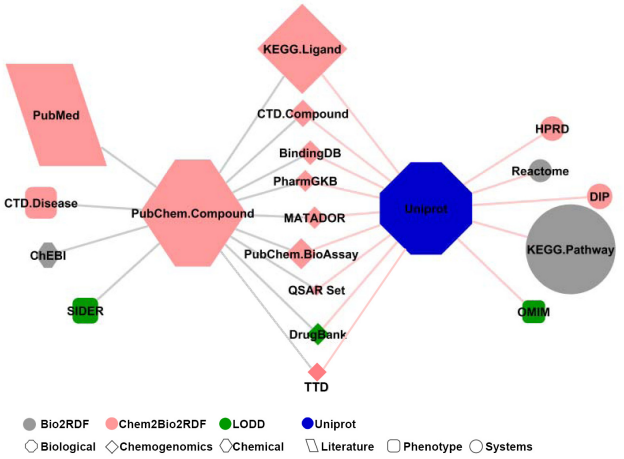

Figure 1: Chem2Bio2RDF for systems chemical biology.

vertically in the middle are the nine major chemogenomics databases [2-10] we mentioned previously. In this view, we remove some cross domain links to accentuate the central role of chemogenomics in bridging the gaps between biology and chemistry hemi-spaces within systems chemical biology space. In this view, we also make Chem2Bio2RDF a comprehensive portal to explore systems chemical biology, for example, one could explicate the adverse drug effect mechanism by tracing a drug reported in SIDER and how it might intervene certain biological pathways through various chemogenomics situation (for example, links through CTD present a comparative toxicogenomics context). Using the graph model, it could be translated into a simple graph search problem that enumerates all the possible paths from *SIDER* to *Pathway* in Figure 1. As it turns out, in total there are 18 such possible schema level links connecting from the instances from SIDER into either Reactome or KEGG (both are public databases dedicated to pathways) ends of Chem2Bio2RDF, since there are nine possible chemogenomics passages to reach over into the systems biology region that either reaches into Reactome or KEGG.

In the technical context of semantic web, the path finding problem in the example above is also referred to as semantic associations, where same pair of instances associated by different linked paths would be ascribed into different contexts, thereby conveying distinctive interpretations. For instance, chemical-gene pairs associated through CTD and PubChem Bioassay would present interaction types attributed to confirmed toxicity cases (CTD) that are distinguishable from ones in a cellular assay (PubChem Bioassay) situation. Once these paths are identified, one can easily translate them into SPARQL queries to retrieve the instance level associations. Given the size of the triple store (80 millions in chem2bio2rdf), the enormity for the amount of instances level associations upon the completion of all 18 queries becomes conceivable. As an example to illustrate the scale here, if we were to find all the chemicals-disease associations related to Alzheimer disease and gene targets involved, at the instance level we found 81077 distinctive Chemical-to-Alzheimer associations that have 410 gene targets involved. There arises an intriguing and important issue – as it is exceedingly laborious if not impossible at all to go through all instance associations, can we devise a good ranking mechanism to present biomedical experts the most important and relevant ones? Or perhaps rare but still significant ones?

In this paper, we develop such a framework to rank semantic associations within the scope of systems chemical biology, and we use chem2bio2rdf as our testbed. In the next section, we relate this work with established formal framework for semantic association.

3. SEMANTIC ASSOCIATION

In semantic web, data objects are connected via properties having formal meanings defined in ontology. Semantic association refers to a sequence of {subject property object} triples that connect two instances, thus the semantic association is contextualized in these properties. For example, Figure 1 illustrates such an association between a chemical compound and a biological pathway. In this case, the chemical compound used as screening agent in PubChem BioAssay (with PubChem compound ID 573747) is linked to the MAPK signaling pathway (with KEGG ID has04010). The data entity and property definitions are provided from the RDF Schema on the left.

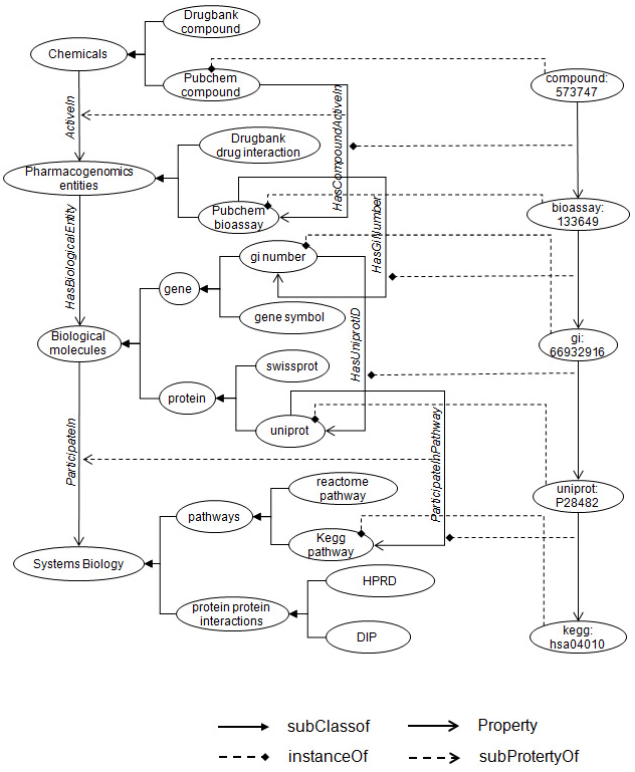

Figure 2. Semantic association between chemical compound and biological pathway.

This association is also consisted of three other intermediary entities and four properties connecting the origin, intermediaries and terminus. Thus the compound-pathway association can be described as the list of triples in the following table:

Table 1. Triple list for compound pathway association

Subject	Predict	Object
Compound:573747	HasCompoundActiveIn	bioassay:133649
bioassay:133649	HasGiNumber	gi:66932916
gi:66932916	HasUniprotID	uniprot:P28482
uniprot:P28482	ParticipateInPathway	kegg:hsa04010

In biomedical research, many important questions can be boiled down to indentifying crucial associations among biological entities. Enclosed in the following list (1-3) are some fundamental questions for systems chemical biology:

1) find gene targets which the compound is active against
2) find compounds associated to a certain disease (Alzheimer)
3) find compounds cause a given adverse drug effect (Hypertension)
4) find all the active compounds in PubChem sharing at least two common targets with a FDA approved drug
5) find all the compounds in PubChem active towards at least two targets that are in the same pathway
6) find KEGG pathways containing at least two of the targets associated with a given side effect

Intuitively, the process of seeking answers for the questions above is equivalent to indentifying entity pairs that satisfy the prescribed associations. Sometimes, answers to the more complicated questions (4-6) could also be synthesized from the simple pair-wise associations, such as finding two association paths joining at a certain instance. In an early framework developed for semantic associations [7], if two instances are connected via two separated paths jointed at a common node, it is then referred as *ρ-join association* and *ρ-path association* if by a directed path (the example in figure 1 would fall under the second category). The following figure presents a specific case of *ρ-join association* in systems chemical biology:

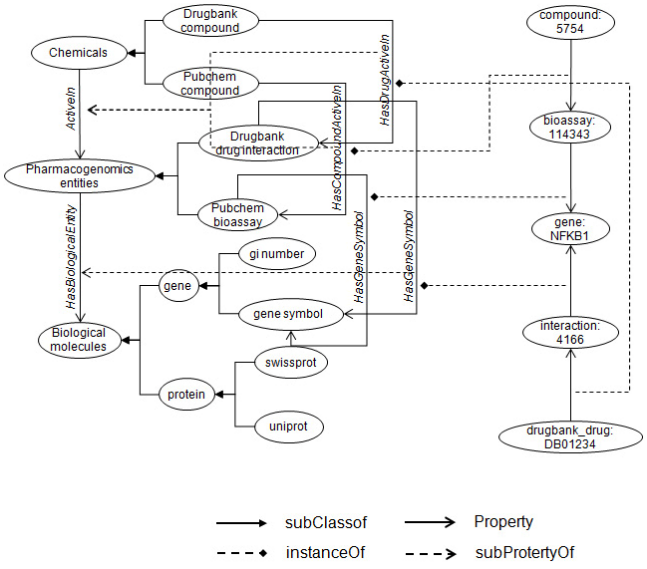

Figure 3. Semantic association between chemical compound and market drug.

In this situation, a compound used in a screening assay (with PubChem compound ID 5754) is associated with Dexamethasone (an anti-inflammatory and immunosuppressant) as they share a common gene target. As it also indicates, the two segment of the joined path (compound to gene, drug to gene) shares instances and properties belongs to the same category i.e. *bioassay* and *drug interaction* are both subclass of *pharmacogenomics entity*; and *HasDrugActiveIn* and *HasCompoundActiveIn* are both subproperty of *ActiveIn*. This satisfies the additional requirement for ρ-join association than ρ-path association specified in [8].

With formal notions introduced above, we can now discuss some issues of particular significance to systems chemical biology. Sometimes in a complicated biological network, there are multiple ways to associate two entities of special interests via different path, be it a path association or join association as we have quoted. The collective views for the set of viable paths sometimes may reveal insights of extraordinary value, and there are two such examples in Figure 4.

Approximately 35% of known drugs or leads have more than one target, and the efficacy of many drugs is increasingly thought to come from their effects on multiple targets. This is known as *polypharmacology*. Of a particular note, if a compound has the same set of multiple targets as a FDA approved drug that compound could be a candidate for a novel therapeutic study. As a) in Figure 4 has shown here, the compound (notice this is the same compound in Figure 3) shares two drug targets of Dexamethasone via two different *ρ-join association*, therefore it may exhibit certain polypharmacology effect as Dexamethasone toward inflammation.

Similarly in b), we found the compound (notice this is the same compound used in Figure 1) is active against two targets in the MAPK signaling pathway via two identical *ρ-path association*. What especially intriguing about this discovery is that, sometimes the drug efficacy get greatly reduced because complicated and redundant pathways often allows an alternative paths to operate after one critical path in the disease-related pathway is blocked by a drug molecule, therefore it is desirable to have a drug molecule active against multiple targets in a pathway to reduce the likelihood of alternative path to compensate the function of the blocked path, and ultimately enhance the drug efficacy.

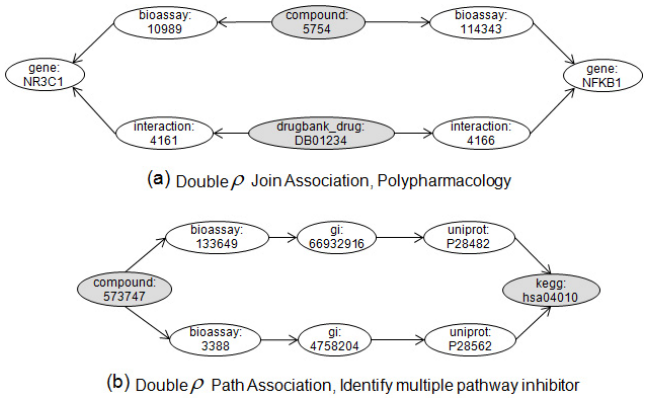

Figure 4. Two intriguing cases for multi-path semantic association .

All the examples in this section demonstrate how natively the intrinsic characteristics and crucial inquiries of systems chemical

biology fit into the framework of semantic association identification. Several types of semantic associations we have introduced in this section reflect different levels of specificity therefore they should have varying weight, i.e. cases conform to polypharmacology and multiple pathway inhibition should be given more attention than the explicit disease-drug or gene-target pair wise associations, therefore a ranking framework is expected. In the next section, we introduce a prototype system to rank semantic associations.

4. METHOD

We have implemented a system called Chem2Bio2RDF Dashboard to support various capabilities for exploring semantic associations for systems chemical biology, the system architecture is shown in Figure 5. The user could formulate queries from a data source ontology aware interface in the format of origin/terminus pairs, between which linked paths are to be identified. The system then executes LPG (Linked Path Generation) – a graph search algorithm to enumerate all simple, non-cyclic paths connecting origin and terminus, all linked paths identified are subsequently converted into the RDF graph patterns used inside the SPARQL queries, which are dispatched to the SPARQL end point at Chem2Bio2RDF portal[1] to retrieve instance level associations. These functions are encapsulated in SPARQL Concretizer module. Figure 6 is a snapshot of it, where totally seven paths are identified, and the third path that connects PubChem BioAssay and Sider via DrugBank translated into the corresponding SPARQL query clauses.

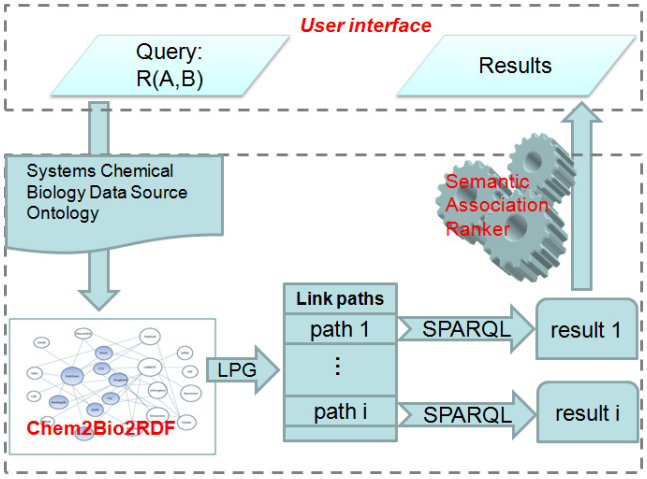

Figure 5. Architecture for Chem2Bio2RDF Dashboard.

At the core of Chem2Bio2RDF Dashboard is the semantic association ranker. As we have introduced previously, the complicated interlinkage between different data sources could easily make tracking associations of interest entities unmanageable. As a case study, we retrieve the outcome for the possible ρ-join associations of compounds and Dexamethasone (the example in Figure 3), as displayed in Table 2. In this case, the retrieved compounds exhibiting activity in their corresponding chemogenomics entities against a gene target, which also participates in a drug interaction that has Dexamethasone involved. The middle column lists the sum of associations found

[1] http://chem2bio2rdf.org/rdf/snorql/

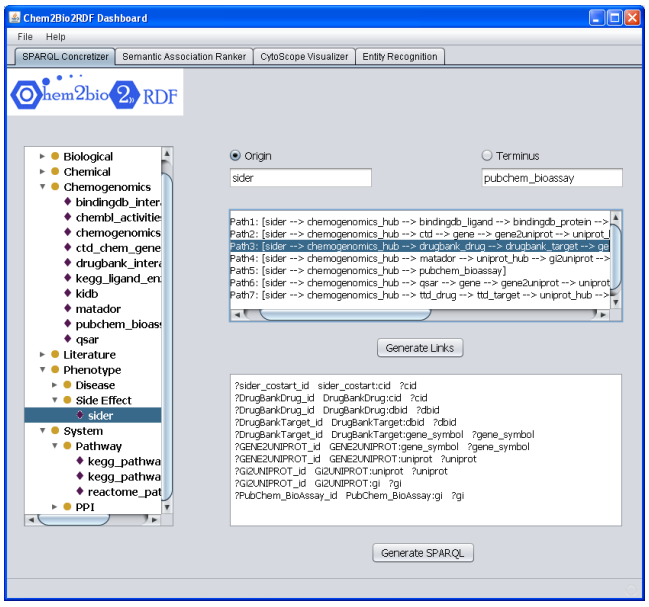

Figure 6. A snapshot for Sparql Concetizer module.

via different chemogenomics data sources. As we can see, even though we have not reconciled the duplicate cases from those data sources, the outcome from PubChem BioAssay along dictates the total amount on a four digit scale. Furthermore, we identify 23 cases exhibiting polypharmacology depicted in Figure 4a).

Table 2. Number of associations returned

data source	number of associations found	polypharmacology cases found
PubChem Bioassay	1123	2
CTD	318	21
BindDB	117	0
TTD	30	0
MATADOR	0	N/A
QSAR	0	N/A
PharmGKB	0	N/A

In order to further arrange the returned associations by their order of significance, we design the following ranking metrics:

Quality: The quality score refers to how reliable a data source is, such that when a semantic association is made through it different trust scores could be assessed. Here we regard manually curated data as the best quality with score of 1, and semi-manually curated and automatically curated with score of 0.5 and 0 respectively.

$$Quality = \begin{cases} 0 & \text{data source automatically curated} \\ 0.5 & \text{data source semi - manually curated} \\ 1 & \text{data source manually curated} \end{cases}$$

Notice here the scoring scheme for quality is designed on an ordinal scale and we make no assumptions that these score assignments reflect the absolute quality. Some of those data sources adopt complete manual curation effort, for instance in

MATADOR, BindingDB and CTD, which will be assigned with quality score of 1 and some adopt semi-manual curation such as in PubChem with score of 0.5.

Specificity: For a biomedical expert, the semantic association paths are valuable not only because they connect entities of interests at the two ends, but also as critical information buried in the intermediary nodes. For example, if a biomedical expert is looking particularly for potent compounds as antagonist, then the associations consist of chemogenomics entities that indicate the participating compound as an antagonist should be weighed above the ordinary ones. Notice that such specific information (antagonist or phosphorylation) is usually coming from the user side and some examples may include antagonist, agonist, metabolic processing, phosphorylation etc. Therefore we design the binary specificity score where it takes on the value of 1 if some intermediary entities within the semantic association contain the specified property and 0 otherwise.

$$Specificity = \begin{cases} 1 & \text{if association path contains specified property} \\ 0 & \text{otherwise} \end{cases}$$

In order to record such scores, the properties for each intermediary entity within an association are checked against the specification (i.e. antagonist or phosphorylation).

Distinctiveness: Some biological entities from the system biology network are effectively hubs that have high connectivity, which lead to frequent appearance within many semantic associations. However, it is not unusual that in biomedical science some rarely occurring instances lead to serendipitous discovery. We design the distinctiveness score here as an indication for high discovery values to less frequent associations.

Table 3. Frequency based distinctiveness score

	CTD	PubChem	TTD	BindingDB	Score
ANXA1	18	0	0	0	0.99
CXCL12	15	0	0	0	0.99
NFKB1	177	989	0	0	0.26
NR3C1	108	134	15	117	0.76

In the above table, the distinctiveness scores based on the gene targets involved in the identified associations are calculated. Each column in table 3 records the number of associations that contain a particular gene target (left most column) and go through a certain data source (header row). Essentially, the distinctiveness score of an certain association with respect to gene target g is treated as the complement of its appearance frequency: $Distinctiveness_g = 1 - Freq_g$, where the denominator in the formula below is the cumulative count for all the identified associations (notice there are 4 gene targets involved in the possible compounds/Dexamethasone ρ-join associations).

$$Freq_g = \frac{count(association_g)}{\sum_{g=1to4} count(association_g)}$$

Multi-Association Accumulator: The individual ranking metrics introduced so far present intuitive assessment for important properties such as quality, trust, rarity and specificity. The linear combination of these scores thus could be deemed as heuristics when ranking the individual semantic associations according to their significance:

RankScore = Quality + Specificity + Distinctiveness

In some special cases, as indicated in the third column of table 2, individual association path that forge double *ρ-join association* (we even found triple and quadruple cases listed in table 4). As we have emphasized in section 3, they are not just simple association paths joined by both of their ends, they actually establish highly valuable cases for polypharmacology study. We thus introduce multi-association accumulator, this operator is applied when we identify same pair of entities that are associated by different paths with exactly same properties and entities of same types, as exemplified in Figure 4. Has this operator applied, individual association paths would be combined into a single case and the scores they carried be accumulated. As the direct result, the ranking score via multiple aggregated paths would be significantly higher than that of a single path, and makes double, triple and quadruple *ρ-join association* cases representing meaningful cases of polypharmacology or multiple pathway inhibitors really standing out in the ranking.

Table 4. Compound (by PubChem ID)/ Dexamethasone ρ-join association ranking outcomes with specificity for phosphorylation

PubChem ID	quality	distinctiveness	specificity	rank score
443495	4	3	0	7
55245	4	3	0	7
969516	3	2.01	0	5.01
74990	3	2.01	0	5.01
5743	2	1.02	1	4.02

As the ranking outcome for the top 5 Compound/Dexamethasone ρ-join association listed in table 4 indicates, the first and the second are quadruple *ρ-join association* that have four distinctive paths relating the respective compounds with Dexamethasone. As we have observed those four association paths each traverse through a distinctive gene in table 3. Also notice that their cumulative distinctiveness scores are effectively sum of the individual distinctiveness scores (sum of the distinctiveness score column in table 3). In table 4, the quality column is the sum of the four individual quality scores, in fact the chemogenomics entities within these associations all come from manually curated sources hence they all take value of 1. The scores for the third and fourth ranked association paths are obtained similarly, except that they are ranked slightly lower because of triple *ρ-join association*. Of a particular note, the fifth association path, given it is a case for double *ρ-join association*, the scores in the first two categories are much lower, neverthess it is gaining one extra point as its chemogenomics entity has a specific property that indicates it matches the phosphorylation specificity requirement whereas the higher ranking cases are void of.

5. CONCLUSION

In this work, we present Chem2Bio2RDF Dashboard – a prototype system for ranking semantic associations within the system chemical biology space. Several important types of semantic association are classified within a formal framework.

We further design ranking metrics to sort semantic association into order of significance.

It should be pointed out here, that the list of ranking metrics is not yet complete, possible addition includes: 1) co-citation score, which considers whether those entities within the same association path appear together in the same scientific literature as an indication of a more concrete and meaningful association type; 2) commonness score, which values the degree of prevalence of a association path, such a score can be seen as the complement for the distinctiveness score. In order to incorporate it into the ranking scheme, we also need to introduce proper coefficient into the ranking scheme.

In the future, we would like to study how well the ranking metrics correlate with expert judgment. Such studies would be instrumental to generate more accurate ranking metrics and provide insights to derive the ones that conform to domain specific interests. We would also like to carry out comprehensive system wide association and ranking studies to further evaluate the efficacy for the method developed in this paper.

6. RELATED WORK

Previous work for generic ranking metrics of semantic associations has been introduced in [9]. The work presented here not only uses such generic metrics but also extends to aggregated paths that reflect specific domain interest (such as polypharmacology). Related works for automatic path generation in semantic web includes RelFinder [10] which automatically searches for semantic paths that connect entities from DBPedia, and render them on interactive interface.

7. ACKNOWLEDGEMENT

This work is funded by NIH VIVO project (UF09179) and Eli Lilly. We want to appreciate School of Informatics and Computing at Indiana University for the support also.

8. REFERENCES

[1] T. I. Oprea, A. Tropsha, J Faulon and M. D. Rintoul. Systems chemical biology. *Nat Chem Biol*, 3:447-450, 2007.

[2] E. K. Neumann. A life science semantic web: are we there yet? *Science*, 283:22–5, 2005.

[3] E. K. Neumann, E. Miller and J. Wilbanks. What the semantic web could do for the life sciences. *Drug Discovery Today:BIOSILICO*, 2:228–34, 2006.

[4] T. Slater, C. Bouton and E. S. Huang. Beyond data integration. *Drug Discovery Today*, 13(13-14):584-9, 2008.

[5] A. Jentzsch, J. Zhao, O. Hassanzadeh, K. Cheung, M. Samwald and B. Andersson. Linking open drug data. Graz, Austria, 2009.

[6] F. Belleau, M. A. Nolin, N. Tourigny, P. Rigault and J. Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41, 706-716, 2008.

[7] K. Anyanwu, A. Sheth, ρ-queries: Enabling Querying for Semantic Associations on the Semantic Web The 12th International World Wide Web Conference, May 2003, Budapest, Hungary, pp. 690-699.

[8] K. Anyanwu, A. Maduko, A. Sheth: "SemRank: Ranking Complex Relationship Search Results on the Semantic Web," In the Proceedings of the 14th International World Wide Web Conference , May 2005, Chiba, Japan, pp. 117-127.

[9] B. Aleman-Meza, C. Halaschek-Wiener, I. Budak Arpinar, C. Ramakrishnan, and A. Sheth, Ranking Complex Relationships on the Semantic Web, IEEE Internet Computing, 9 (3), May-June, pp. 37-44, 2005.

[10] P. Heim, S Hellmann, J. Lehmann, S. Lohmann and T. Stegemann. RelFinder: Revealing Relationships in RDF Knowledge Bases. SAMT, Lecture Notes in Computer Science, 5887: 182-7, 2009.