The knowledge-driven exploration of integrated biomedical knowledge sources facilitates the generation of new hypotheses

Vinh Nguyen[1], Olivier Bodenreider[2], Todd Minning[3], and Amit Sheth[1]

[1] Kno.e.sis Center, Wright State University, Dayton, Ohio
[2] National Library of Medicine, Bethesda, Maryland
[3] Center for Tropical and Emerging Global Diseases, Univeristy of Georgia, Georgia

Abstract. Knowledge gained from the scientific literature can complement newly obtained experimental data in helping researchers understand the pathological processes underlying diseases. However, unless the scientific literature and experimental data are semantically integrated, it is generally difficult for scientists to exploit the two sources effectively. We argue that, in addition to the semantic integration of heterogeneous knowledge sources, the usability of the integrated resource by scientists is dependent upon the availability of knowledge visualization and exploration tools. Moreover, the integration techniques must be scalable and the exploration interfaces must be easy to use by bench scientists. The end goal of such integrated knowledge sources and exploration tools is to enable scientists to generate novel hypotheses from the knowledge they explore. We tested the feasibility of our approach on a real use case in the domain of human health and parasite biology. On the one hand, we integrated the experimental data generated as part of an ongoing research on Chagas disease with the knowledge extracted from the PubMed articles, using Semantic Web technologies. On the other hand, we developed iExplore, a web tool with a graphical interface for interactive knowledge exploration, that allows non-technical users to explore the integrated knowledge base using a relationship-focused approach. We illustrate the effectiveness of our approach by describing the knowledge-driven process of using iExplore to generate a new hypothesis for the treatment of Chagas disease.

1 Introduction

Translating the knowledge from basic research into practice has been an important trend in biomedical research in recent years. Translational research aims to improve health by utilizing a wide range of biomedical resources, using knowledge from experimental data at the point of care and guiding basic research with problems encountered in patients. A large amount of biomedical knowledge is available in the biomedical literature, e.g., in PubMed[1] articles, and in structured

[1] http://www.ncbi.nlm.nih.gov/pubmed

knowledge sources, including the Unified Medical Language System (UMLS) [2] and the Entrez Gene database. Recent research [4, 6] has shown the potential of using text mining on PubMed articles to advance the biomedical research by exploiting the associations of genes and diseases from the text to prioritize the candidate genes. However, we believe that these approaches do not exploit their potential fully, because their context is limited by design to a specific subset of the biomedical literature.

We argue that a broad knowledge base is key to enabling the generation of new hypotheses. Using a large subset of the scientific literature will benefit biomedical scientists performing basic or clinical research, and facilitate the integration of their data with the biomedical literature. We anchored our knowledge base in comprehensive resources, such as the UMLS [2] and Entrez Gene. The UMLS Metathesaurus enables the interoperability among data sources by providing reference identifiers for biomedical entities. Moreover, the UMLS Semantic Network provides a consistent categorization of the UMLS Metathesaurus concepts, together with a set of relations among the categories. The Entrez Gene database contains the list of genes for various model organisms, as well as their annotations. The combination of UMLS and Entrez Gene in the schema of the Biomedical Knowledge Repository (BKR) provides the flexibility and scalability for integrating additional data sources. We illustrate the integration of the BKR with the experimental data obtained from the research in Chagas disease in section 2.

We believe that the knowledge exploration process should be supported by a tool that the scientists find easy to use and effective to help them gain new insights from the integrated knowledge bases of text and experimental data. The visualization should display sufficient contextual information to the users, and the navigation should be driven by the intuition and background knowledge of the scientists. We developed iExplore, a web tool that displays the graph from integrated knowledge bases in an interactive manner using a relationship-centric approach. The tool complements to the function of existing tools, e.g., RelFinder[2]. While iExplore supports the exploration process, it is not supposed to replace the role of the scientists in this process. We explain the knowledge exploration process enabled by this tool in section 3.

2 Integration of knowledge sources

2.1 The Biomedical Knowledge Repository

The Biomedical Knowledge Repository (BKR) aims to integrate knowledge from a variety of sources ranging from the scientific literature to various structured knowledge bases [3]. The BKR contains relations extracted from PubMed documents by SemRep[1] and normalizes biological entities to concepts in the UMLS and Entrez Genes. It includes approximately 20 million semantic predications

[2] http://www.visualdataweb.org/relfinder.php

extracted from 6 million articles in PubMed published from 1999 to 2009. These semantic predications are transformed into RDF format together with the provenance information about the article where the predication is extracted.

Table 1. RDF triples represent "CALR associated_with Chagas Disease"

Triple	Subject	Predicate	Object
1	META_C0041234	rdfs:label	"Chagas Disease"
2	EG_811	rdfs:label	"CALR"
3	META_C0041234_INST	rdf:type	META_C0041234
4	PUBMED_19108895/EG_811	rdf:type	EG_811
5	PUBMED_19108895/EG_811	associated_with	META_C0041234-INST
6	PUBMED_19108895/EG_811	derives_from	PUBMED_19108895

A semantic predication extracted from the title or abstract of an article is represented as a set of RDF triples. For example, the title *"Trypanosoma cruzi calreticulin: a possible role in Chagas' disease autoimmunity"* of the article with PubMed ID 19108895 contains one predication, "CALR associated_with Chagas Disease". The set of triples in Table 1 is created to represent this semantic predication in the BKR.

2.2 Experimental Data about Chagas Disease

The experimental data about Chagas Disease originate from DNA microarray analysis, proteome analysis, gene knockout and strain creation protocols. DNA microarray analysis was used to measure the relative transcript abundances for all of the genes in the *T. cruzi* genome during the four main life cycle stages of *T. cruzi*, namely amastigote, trypomastigote, epimastigote and metacylic trypomastigote. Whole genome shot-gun proteomic analysis was used to measure the presence of proteins encoded by *T. cruzi* genes during the four life cycle stages. The proteome and transcriptome data have been used to prioritize genes for the gene knockout and strain creation protocols. To capture the detail of these experimental protocols, we created two ontolologies: Parasite Experiment and Parasite Lifecycle, that have been published in the BioPortal[3]. We use these ontologies as schema to convert the experimental data into RDF.

2.3 Integration

A unified view of experimental data with biomedical literature requires the mappings between entities of two data sources. Genes are one of the potential common entities between the BKR and experimental data sources. However, the experimental data are based on the *Trypanosoma cruzi (T. cruzi)*

[3] http://bioportal.bioontology.org/

genome, while predications extracted from the biomedical literature refer to human genes. The study of the human orthologs of *T. cruzi* genes is helpful because the gene function is usually conserved across species. To bridge the gap between genes of two organisms, we use the orthologous mapping from *T. cruzi* to *Homo sapiens*(human) from the KEGG Sequence Similarity database[4], and create an RDF triple for each pair of orthologous genes. For example, orthology between the human gene "CALR" (ID 811 in Entrez Gene) and the *T. cruzi* gene Tc00.1047053509011.40 is represented by the following RDF triple: "EG_811 is_orthologous Tc00.1047053509011.40." In practice, such orthology relations connect entities across the two sources.

3 Knowledge exploration

Knowledge exploration is the process of establishing a new relationship between two known concepts and generalizes the notion of literature-based discovery to sources others than the biomedical literature [5]. The exploration process includes two steps, navigation and interpretation.

Navigation requires the visualization of knowledge as a set of relations. Such relations can be explored by the user based on domain knowledge. We developed iExplore[5] with an intuitive graphical interface to enable navigation. The tool creates an abstraction layer from the RDF integrated knowledge bases to visualize the graph of concepts and named relationships. Because knowledge is represented as a set of relations, users can expand and narrow the graph based on their interest. Graph expansion is implemented through predefined SPARQL queries hidden from users. In practice, biologists drive the exploration guided by their background knowledge, selecting concepts to expand the graph or restricting the graph to specific relations. **Interpretation** allows biologists to utilize their background knowledge to generate novel hypotheses from the chains of concepts identified in the navigation phase.

Example The exploration starts by expanding the concept "Chagas Disease" and inspecting its related concepts. Of particular interest to us are known treatments for Chagas disease. We restrict the graph to the "TREATS" relation using a filter. Among the treatment concepts, we focus on drugs (categorized by semantic types, such as "Pharmacologic Substance"). We find the drug itraconazole, known for treating various parasitic diseases and which has side effects. We pursue our exploration by expanding the graph with the relations of the concept "Itraconazole". Specifically, we want to explore the genes connected to "Itraconazole" via the "INHIBITS" relation because they possibly indicate biological pathways involved in the treatment of Chagas disease. Since only human genes are present in the graph extracted from the literature, we follow the orthology relation in order to find the *T. cruzi* orthologs of these human genes, i.e., the

[4] http://www.genome.jp/kegg/ssdb/

[5] Due to limited space, the tool and illustrated examples in section 3 are presented in the tool's homepage http://knoesis.wright.edu/iExplore/index.html for review.

possible target of the drug in the parasite. In summary, this example establishes a chain of named relationships from "Chagas Disease" to itraconazole and human genes to *T. cruzi* genes. We generate a hypothesis from this chain, that the *T. cruzi* orthologs of human genes inhibited by itraconazole may also be inhibited by itraconazole and thus would be candidates for further studies into the mechanism(s) of action of the drug itraconazole on *T. cruzi*.

4 Conclusion

We presented the semantic approach that underpins the integration and complementary use of the two independently generated, heterogeneous knowledge sources. These integrated knowledge bases together with iExplore allow biologists to use their knowledge to drive the exploration and generate new biomedical hypotheses. The validation of such hypotheses is part of our future work. We also plan to improve iExplore by learning the way biologists use their background knowledge to generate hypothesis, and then automate the interpretation step by making recommendations for hypothesis generation. Of note, our approach can easily be generalized to other diseases and, more generally, to other data sources integrated and explored together with the BKR following the approach we demonstrated with Chagas disease. The two-step exploration process can also be applied to make use of the broader integration of these independently generated knowledge sources.

Acknowledgements This research was supported by an appointment to the Research Participation Program at National Library of Medicine, and the NIH R01 Grant number 1R01HL087795-01A1. We also acknowledge Dr. Thomas Rindflesch, Dr. Cartic Ramakrishnan, Dr. Priti Parikh, Jonathan Mortensen, Joshua Dotson and Sarasi Lalithsena for help.

References

1. C. Ahlers, M. Fiszman, D. Demner-Fushman, F. Lang, and T. Rindflesch. Extracting semantic predications from medline citations for pharmacogenomics. In *Pacific Symposium on Biocomputing*, 2006.
2. O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32, 2004.
3. O. Bodenreider and T. Rindflesch. Advanced library services: Developing a biomedical knowledge repository to support advanced information management applications. *National Library of Medicine*, 2006.
4. A. Faro, D. Giordano, and C. Spampinato. Combining literature text mining with microarray data: advances for system biology modeling. *Briefings in Bioinformatics*, 2011.
5. D. Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7, 1986.
6. N. Tiffin, J. Kelso, A. Powell, H. Pan, V. Bajic, and W. Hide. Integration of text-and data-mining using ontologies successfully selects disease gene candidates. *Nucleic acids research*, 33(5):1544, 2005.