From Questions to Effective Answers: On the Utility of Knowledge-Driven Querying Systems for Life Sciences Data

Amir H. Asiaee[1], Prashant Doshi[1], Todd Minning[2], Satya Sahoo[3], Priti Parikh[3], Amit Sheth[3] and Rick L. Tarleton[2]

[1] THINC Lab, Dept. of Computer Science, University of Georgia, Athens, GA
[2] Tarleton Research Group, Dept. of Cellular Biology, University of Georgia, Athens, GA
[3] Kno.e.sis Center, Dept. of Computer Science, Wright State University, Dayton, OH
aha@uga.edu, pdoshi@cs.uga.edu, {tminning,tarleton}@uga.edu, {satya,priti,amit}@knoesis.org

Abstract. We compare two distinct approaches for querying data in the context of the life sciences. The first approach utilizes conventional databases to store the data and provides intuitive form-based interfaces to facilitate querying of the data, commonly used by the life science researchers that we study. The second approach utilizes a large OWL ontology and the same datasets associated as RDF instances of the ontology. Both approaches are being used in parallel by a team of cell biologists in their daily research activities, with the objective of gradually replacing the conventional approach with the knowledge-driven one. We describe several benefits of the knowledge-driven approach in comparison to the traditional one, and highlight a few limitations. We believe that our analysis not only explicitly highlights the benefits and limitations of semantic Web technologies in the context of life sciences but also contributes toward effective ways of translating a question in a researcher's mind into precise queries with the intent of obtaining effective answers.

1 Introduction

Much of the data in the life sciences continues to be stored using conventional database management systems (DBMS) and subsequently, queried using the structured query language (SQL). Intuitive interfaces such as forms often provide and support "pre-canned" queries that are most commonly used by the researchers who are chiefly interested in quick and targeted accessibility to the data. However, these interfaces tend to provide more data than needed leading to time-consuming post processing steps which are specific to the local researchers, instead of being general.

We compare and contrast two approaches for querying life sciences data. Both utilize an identical data context: *strain*, *stage transcriptome* and *proteomic* data on the parasite *Trypanosoma cruzi* (*T. cruzi*). In the first approach, *T. cruzi* data is stored in a conventional DBMS and accessed through a suite of well-designed forms representing a predefined set of queries, we refer to this approach as Paige Tools [1] which has

been the de-facto way for storing and accessing experimental data related to *T. cruzi* by the Center for Tropical and Emerging Diseases at the University of Georgia. The second approach, *Parasite Knowledge Repository* - PKR, uses an OWL-based ontology designed in collaboration with the life science researchers to model *T. cruzi* experimental data [2]. Querying capabilities of PKR are provided by an enhanced version of a knowledge-driven querying system, *Cuebee* [3] [4], that facilitates formulation of RDF triple-based queries, which are transformed to SPARQL-DL [5].

We believe that Paige Tools and PKR is representative of the traditional and more sophisticated way of querying life sciences data, respectively. These approaches provide alternative ways of transforming the precise question in a researcher's mind into a computational query in order to obtain the answer. The outcome of our analysis is a set of benefits that knowledge-driven approaches such as PKR offer over the more conventional approaches. We also highlight two limitations that this approach faces, which could impede its widespread adoption despite the substantial benefits.

2 Related Work

Other Semantic Web based systems exist that focus on queries to provide targeted access to data in the life sciences and other contexts. These include query tools such as Openlink iSPARQL [6] and NITELIGHT [7] both of which provide graph-based interfaces for query formulation. These systems did not provide evaluation of their approaches on real-world data. Similar to PKR, GINSENG [8] offers suggestions to users, but from a different perspective. GINSENG relies on a simple question grammar, which is extended using the ontology schema to guide users to directly formulate SPARQL queries. Bernstein et al. [8] briefly evaluated GINSENG on three aspects: usability of the system in a realistic task, ability to parse large number of real-world queries, and query performance.

Semantics-based approaches also exist that focus more on data integration in the life sciences context. GoWeb [9] is a semantic search engine for the life sciences, which combines keyword-based Web search with text-mining and ontologies to facilitate question answering. GoWeb demonstrates a recall of 55 to 79% on three benchmarks. Cheung et al. [10] introduce semantic Web query federation in the context of neuroscience which provides facilities to integrate different data sources and offers either SPARQL or SQL query. Mendes et al. [4] evaluated the *usability* of *Cuebee* on the system usability scale [11] and the query formulation effort by recording time taken and number of interactions to retrieve answers. Because PKR's front end uses an enhanced version of *Cuebee* we believe that the same evaluation holds.

All of the listed approaches are available for public use. However, there is not enough evidence of how much these systems are in use by life science researchers in daily research. This paper discusses significant enhancements to *Cuebee* [3] [4], and explicitly highlights the benefits and limitations of using PKR while being used by an interdisciplinary team of computer science and cell biology researchers. Thus, while PKR is not alone in bringing knowledge-driven approaches to the life sciences, we believe that our comparative evaluation of the systems in use is novel.

3 Background

In this section, we briefly describe the two approaches for querying experimental data related to *T. cruzi*. We emphasize that both Paige Tools and PKR are currently operational and are being used by researchers, with the expected longer-term objective of replacing Paige Tools with PKR.

3.1 Paige Tools – Conventional DBMS-based Approach

Paige Tools offers interfaces to add and edit experimental data related to *T. cruzi* housed in multiple separate local databases as well as facilities to execute queries. Typically, these interfaces manifest as forms containing widgets such as drop-down lists, check boxes and buttons that allow formulation of a Boolean query on a specific dataset and selection of attributes to display in the result. We believe that the interfaces in Paige Tools are typical of systems utilized by life science researchers. As expressed by the researchers that use Paige Tools, these tend to be simple but adequate approaches for somewhat targeted access to portions of data. The interfaces are tightly coupled to the schema design and limited to executing a specific set of queries. Thus, any change to the database schema results in refactoring of the forms.

3.2 PKR – Knowledge-Driven Approach

At the front end of PKR we use a significantly enhanced version of *Cuebee* – an ontology-based query formulation and data retrieval system applied in the context of *T. cruzi* parasite research originally designed by Mendes et al. [3] [4].

Cuebee employs two query engines, which we refer to as *suggestion engine* and *answer engine*. *Suggestion engine* guides a user through the process of transforming her question into a query in a logical way. It utilizes RDFS ontology schemas to suggest concepts in a drop-down list that match the characters that the user types. Furthermore, it lists all the relevant relationships for any selected particular concept. In the process of formulating the query users may need to select some intermediate concepts in order to relate the concepts that appear in the question. Finally, queries are transformed into SPARQL queries and executed by the *answer engine*.

We introduce multiple enhancements to make *Cuebee* more user-friendly [12]. For example, the enhanced *suggestion engine* now annotates each suggested concept with information that includes a description of the ontology class and associated properties. It allows selection of multiple instances that satisfy Boolean operators. The enhanced *Cuebee* also guides users to formulate more complex SPARQL graph patterns using group by and aggregate functions, filter over instances using regular expressions. In addition, an undo feature helps users revise their queries at any point during the formulation process.

Our contributions go beyond the interface and focus on the infrastructure of *Cuebee* as well. A major improvement is the capability to support OWL ontologies because they tend to be more expressive than RDFS ontologies. For example, in the

context of *T. cruzi* research, we use the OWL-based parasite experiment (PEO) and parasite lifecycle (OPL) ontologies [2]. Subsequently, we equip the two query engines to execute SPARQL-DL [5] queries which offer more expressive power than SPARQL. OWL ontologies are deployed in an OWL-DL reasoner called *Pellet* in order to take advantage of the inferencing capabilities.

An increasing number of bioinformatics tools and biomedical data sources are available as Web services. As another contribution to *Cuebee*, we extend the results of the final queries with common bioinformatics tools such as EBI BLAST available as RESTful Web services and access into TriTrypDB [13]. Here, we detect if the results of a query contain appropriate types of protein sequences or gene IDs, and allow the user to trigger an invocation of the EBI BLAST Web service or obtain additional information from TriTrypDB.

4 Benefits of PKR over Paige Tools

Both Paige Tools and PKR are running concurrently on identical data and in use by a team of researchers. The identical contexts provide us a valuable opportunity to comparatively evaluate the two approaches in a principled way in this section.

4.1 Explicitly Structured Queries

The first benefit is with respect to the structure of the queries that may be formulated in the two approaches. In order to illustrate this, consider the following question posed by parasitologists in the context of *T. cruzi*:

Which microarray oligonucleotide derived from homologous genes has 3 prime region primers?

Note that homology is a relationship between two genes (these genes are derived from a common ancestor) and *3-prime-region* is a property of primers.

Conventional database design places minimal importance on named relationships (e.g., table joins) and Paige Tools as a typical example of DBMS-based systems that are in use in life science research labs, reflects this. While query pages within Paige Tools provide users the ability to show attributes of *microarray oligonucleotide*, *genes* and *primers*, discerning homology relationships between two genes is left to the ability of the user in post-processing the results. Thus, the resulting query does not adequately reflect the original question in the researcher's mind.

Figure 1. Formulated query for "*Which microarray oligonucleotide derived from homologous genes has 3 prime region primers?*" in PKR. Notice the relationships between the concepts.

On the other hand, PKR's process of formulating queries allows a logical interpretation of the question. Queries formulated within PKR contain not only the concepts (e.g., *gene*) but also make the relationships explicit in the query (e.g., *is*

homologous to), as we show in Fig. 1. The query formulation process in PKR leads users to find linkages between concepts by suggesting relationships explicitly. Due to the expressiveness of ontology schemas, the formulated query is more readable and promotes better understanding even to users with less domain knowledge.

4.2 Queries at Different Levels of Abstraction

A significant benefit of PKR is its ability to query at multiple levels of abstraction. This is beneficial because researchers investigating new hypotheses often ask general questions. Consider the following question posed by our parasite researchers:

What genes are used to create any T. cruzi sample?

T. cruzi sample could be of several different types: *cloned sample*, *drug selected sample*, and *transfected sample*. There is no straightforward way to transform this general question into a query using Paige Tools. A researcher translates this question into a query for strains database that produces almost all genomic data. Then, the researcher tediously analyzes multiple attributes for each data record to ascertain the type of *T. cruzi sample*. In this approach, explicitly linking the different samples would involve redesigning the database and reduced efficiency.

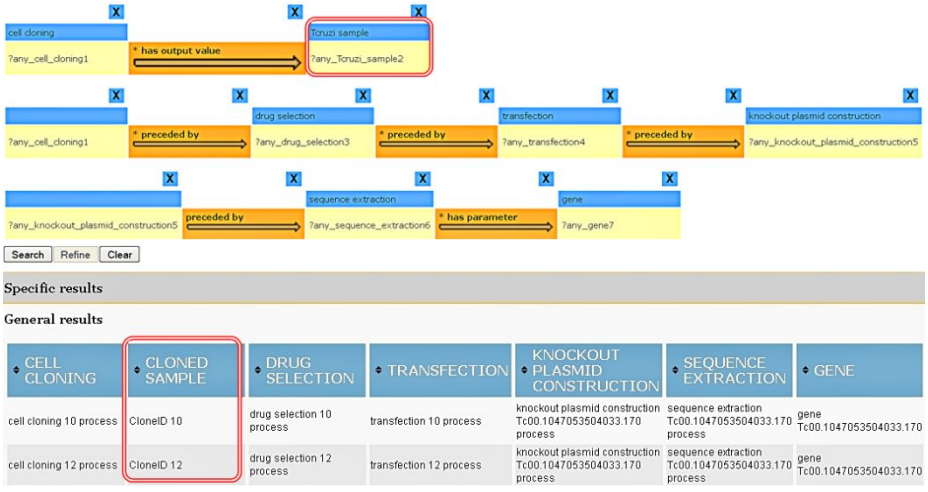

Figure 2. The question "*What genes are used to create any T. cruzi sample*?" is formulated in PKR and *cloned sample* which is a type of *T. cruzi sample* appears in the results.

On the other hand, PKR intuitively models the relationships between the different types of samples in the ontology schema. PKR's *answer engine* takes advantage of Pellet's inferencing by using SPARQL-DL's extended vocabulary and generates the corresponding query in order to access instances of the class and all its subclasses. As Fig. 2 illustrates, *cloned sample* – a subclass of *T. cruzi sample* – appears under the "General Results" tab. Therefore, answering general questions is less dependent on a user's domain expertise in contrast to Paige Tools.

4.3 Uniform Query Interface

Ontology-driven approaches such as PKR allow a uniform query interface for multiple related datasets; however, Paige Tools offers several interfaces to access the different databases. Each interface is designed using drop-down lists holding different attribute names from the corresponding table schema and check boxes to give the option to the user of filtering results (see Fig. 3). Notice that the items in the drop-down lists and the check box labels differ across the two interfaces.

PKR provides a uniform query interface to the user regardless of which datasets are the target of the questions. The process of translating the question into a query does not change with different contexts. By default, formulated queries are executed over all datasets. Users may also select a suitable dataset from the drop-down list of datasets for efficiency. This is enabled by using a single, comprehensive ontology schema for the related datasets. Furthermore, approaches such as PKR are usually not tied to a specific ontology but support any ontology designed in OWL.

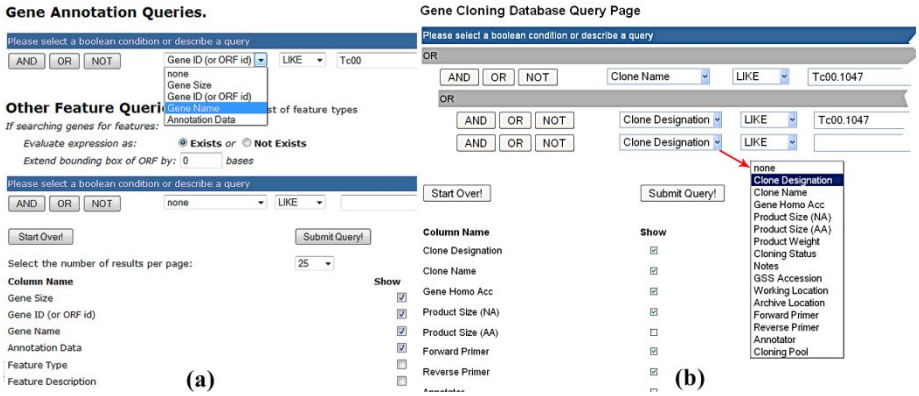

Figure 3. The (a) gene annotation query page and (b) cloning database query page – representing two interfaces of Paige Tools.

4.4 Querying over Multiple Datasets

Often, researchers pose questions that span across different types of data. For example, consider the following question:

Which genes with log-base-2-ratio greater than 1 have 3 prime region primers?

Data related to *log-base2-ratio* is found in the *transcriptome* dataset while primers with *3-prime-regions* are found in *strain* dataset. In Paige Tools question is divided into two sequential sub-questions: (a) *Which genes have log-base-2-ratio greater than 1*; and (b) *which of these genes has 3-prime-region primers*. Answer to question (a) is found using the gene annotations query page. Then, a researcher takes the results from (a) and manually looks for the primers in the gene cloning page to find answers to (b).

On the other hand, PKR allows a formulation of the associated query without decomposing it despite the fact that two different datasets hold the answers. A user finds the appropriate concepts and relationships between *log-base-2-ratio* and *gene* (Fig. 4

area (1)), and continues to formulate the query by adding the *has-3-prime-region* relationship followed by region (Fig. 4 area (2)). On formulating the query, PKR allows a search over all datasets – made possible because of a comprehensive ontology for all the data. The solution to the query integrates both datasets thereby facilitating integrated analysis by the researchers with minimal post-processing effort.

Figure 4. The question, "*Which genes with log-base-2-ratio greater than 1 have 3 prime region primers*", formulated in PKR. The query for this question spans multiple datasets.

5 Limitations of PKR

We highlight two limitations of approaches such as PKR, which may likely impact its widespread adoption. While ontologies represent a formal model of the domain knowledge, users not well acquainted with the ontology feel tied down to its structure. We minimize this by providing suggestions about next possible concepts and relationships. Nevertheless, our triple-based queries often require users to select intermediate concepts and relationships that connect the entities in the question. But users prefer more abbreviated queries in their daily usage of systems such as PKR.

The second limitation is the increased time and space complexity of knowledge-driven systems compared to highly optimized modern DBMS. While fast RDF storages such as *Virtuoso* exist, the predominant complexity is due to the ontology inferencing facilities provided by systems such as *Pellet.*

6 Evaluation and Discussion

While Mendes et al. [4] evaluated the usability of PKR's interface, in this paper, we focus on the *usefulness* of knowledge-driven systems such as PKR in comparison to DBMS-based systems such as Paige Tools, which requires that the systems be in use. We compile our observations of both systems in use into the benefits and limitations of the two approaches, in Sections 4 and 5. In order to quantify aspects of usefulness of PKR and Paige Tools we calculate precision and recall on a corpus of 25 domain questions, many of which span multiple datasets. Although the domain of these questions is limited to the parasite, *T. cruzi*, such questions are commonly encountered by biologists and parasitologists investigating other organisms as well.

Two domain experts independently validated the consensual reference set for each question in this evaluation. We obtain average precisions of 83% and 56% for PKR and Paige Tools, respectively; average recall score for PKR is 80% and for Paige Tools is 77%. Our results show that both systems retrieve large fractions of the rele-

vant data from the collection of all data, and queries in PKR provide more accurate answers than in Paige Tools. The latter lead to much post processing, as mentioned.

Parasitologists using PKR appreciate its advantages and are getting more comfortable with the layout as it improves. But, it takes time to get researchers to change over completely. We are not yet at a point where researchers in other labs may be able to simply install PKR and query their particular sets of data. Many of the concepts used in PEO are general enough to be incorporated into ontologies for other organisms, but we anticipate that ontologies will still require tailoring to individual use cases. The scope of this paper is to provide a model for developing ontology-based systems for life science researchers, to offer proof that semantic Web technologies will ultimately be of greater use to biomedical researchers than traditional DBMS, and to demonstrate the capabilities of PKR. We believe that these are substantive steps towards developing systems that are more user friendly and efficient for biomedical researchers. As PKR continues to be utilized we expect that researchers will gain new biological insights from their analysis of the data.

7 Reference

1. Paige tools. In: . Available at: http://paige.ctegd.uga.edu
2. Parikh, P., Minning, T., Nguyen, V., Lalithsena, S., Asiaee, A., Sahoo, S., Doshi, P., Tarleton, R., Sheth, A.: A Semantic Problem Solving Environment for Integrative Parasite Research: Identification of Intervention Targets for Trypanosoma cruzi. PLoS Neglected Tropical Diseases 6(1)(e1458) (2012)
3. Cuebee (Original Version). Available for download at: http://Cuebee.sourceforge.net
4. Mendes, P., McKnight, B., Sheth, A., Kissinger, J.: Tcruzikb: Enabling complex queries for genomic data exploration. In : IEEE-ICSC, pp.432-439 (2008)
5. Sirin, E., Parsia, B.: SPARQL-DL: SPARQL Query for OWL-DL. In : Third OWL Experiences and Directions Workshop (OWLED) (2007)
6. Kiefer, C., Bernstein, A., Lee, H., Klein, M., Stocker, M.: Semantic Process Retrieval with iSPARQL. In : ESWC Springer-Verlag, Berlin, pp.609-623 (2007)
7. Russell, A., Smart, P., Braines, D., Shadbolt, N.: NITELIGHT: A Graphical Tool for Semantic Query Construction. In : SWUI hosted by CHI, Florence (2008)
8. Bernstein, A., Kaufmann, E., Kaiser, C., Kiefer, C.: Ginseng: A Guided Input Natural Language Search Engine for Querying Ontologies. In : Jena User Conference, UK (2006)
9. Dietze, H., Schroeder, M.: GoWeb: a semantic search engine for the life science web. BMC Bioinformatics 10(Suppl 10), S:7 (2009)
10. Cheung, K., Frost, H., Marshall, M., Prud'hommeaux, E., Samwald, M., Zhao, J., Paschke, A.: A journey to Semantic Web query federation in the life sciences. BMC Bioinformatics 10(Suppl 10), S:10 (2009)
11. J., B.: SUS: a quick and dirty usability scale. In : Usability evaluation in industry pp.189–194. (1996)
12. Cuebee (Enhanced Version). Available at: http://jade.cs.uga.edu:8080/Cuebee
13. TriTrypDB. Available at: http://tritrypdb.org