A large-scale gene-centric semantic web knowledge base for molecular biology

José Cruz-Toledo[1§], Alison Callahan[1§], and Michel Dumontier[1]

[1]Department of Biology, Carleton University, Ottawa, Canada,
{acallaha, jctoledo}@connect.carleton.ca, michel_dumontier@carleton.ca
[§]These authors contributed equally to this work

Abstract. The discovery of the central role of genes in regulating the fundamental biochemical processes of living things has driven biologists to collect, analyze and re-use enormous amounts of information, and to make this information available in thousands of curated databases. The increasingly popular use of specialized terminologies, often organized into hierarchical taxonomies or more formal ontologies, to describe this data indicates that managing the total amount of resources available (big data) will surely continue to be an ongoing challenge. Here, we describe a biological gene-centric dataset (available at http://semanticscience.org/projects/gene-world), aimed at providing the reasoner community with a fully connected graph of data and ontologies of value to the bioinformatics community and for which there currently exists significant challenges in using automated reasoning for consistency checking and query answering of large ontology-mapped linked data.

Keywords. Semantic Web, Bioinformatics, DL reasoning, SPARQL

1 Motivation

The central dogma of molecular biology states that regions of DNA (genes) are responsible for encoding molecular machines called proteins, which participate in and control the biochemical reactions essential to sustaining life. The discovery of the central role of genes as the blueprint of our evolutionary history and their involvement in health and disease has driven biologists to characterize these very important entities. Enormous amounts of information have been collected, analyzed, summarized and re-published in thousands of curated databases [1] and large central hubs such as the databases of the National Center for Biotechnology Information (NCBI).

The exponential growth of available molecular data clearly yields enormous benefits to biologists attempting to elucidate the functioning of genes in related systems, but it also presents significant challenges for modern biology. Consider that the amount of data collected in this year alone to characterize a collection of biochemical reactions (a pathway) will be on par with the amount of data that has ever been collected about that pathway in the history of the field [2]. Moreover, the use of specialized terminologies, often organized into hierarchical taxonomies or more formal ontologies, indicates that managing the total amount of data (big data) will surely con-

tinue to be an ongoing challenge. Indeed, it is the overall organization and interpretation of this vast deluge of information that presents the greatest challenge.

Motivated by this challenge, we present a preliminary version of a biological gene-centric dataset aimed at providing the reasoner community with a fully connected graph of data and ontologies of value to the bioinformatics community, for which there currently exists significant challenges in using automated reasoning for consistency checking and query answering of large ontology-mapped linked data. We focus our attention on one of the larger datasets in the Bio2RDF project [3] - NCBI Gene - and consider queries that extend from this dataset into other datasets and ontologies that together form a large 'Gene-World' knowledge base. Our Gene-World knowledge base contrasts other ontologies and datasets that have been used to benchmark OWL reasoners [4], such as LUBM [5] and SNOMED-CT [6], in several respects: (i) Gene-World is a 'real world' knowledge base composed of existing resources used by biologists and bioinformaticians on a daily basis, as opposed to an arbitrary automatically generated knowledge base, (ii) it consists of a very large T-box and A-box and (iii) its T-Box consists of multiple ontologies with differing DL expressivity. The datasets and ontologies described are available at [7]. Example queries that can be used to evaluate RDF/OWL based reasoner performance over this knowledge base are also described.

2 Datasets and Ontologies

All Gene-World datasets are drawn from Bio2RDF Release 2 (released January 2013). The NCBI Gene [8] Bio2RDF dataset consists of 394,026,267 triples with 12,543,449 unique subjects, 60 unique predicates, and 121,538,103 unique objects. NCBI Gene describes genes including their names, reference sequences, variants, phenotypes, pathways and cross-references to related resources. HomoloGene [9] is a database of programmatically generated clusters of homologous, including paralogous and orthologous, genes from a set of 21 completely sequenced eukaryotic genomes. The HomoloGene Bio2RDF dataset consists of 1,281,881 triples with 43,605 unique subjects, 17 unique predicates and 1,011,783 unique objects, and uses NCBI Gene identifiers to refer to the genes it clusters. NCBI Gene makes reference to three ontologies: the Gene Ontology (GO) for asserting function, process or location annotations about genes, the Evidence Code Ontology (ECO) for qualifying the source of these GO annotations, the NCBI Taxonomy (TAXON) for asserting the species of a gene. The Semanticscience Integrated Ontology (SIO) and the Sequence Ontology (SO) have been mapped to NCBI Gene Bio2RDF vocabulary classes and relations (Table 1) to ground the dataset types and predicates in domain-specific ontologies.

The Gene Ontology (GO) [10] [11] is a hierarchy of controlled biological terms that is organized into three orthogonal ontologies which capture knowledge about cellular locations, biological processes and molecular functions. The terms and relations contained in GO are serialized as a directed acyclic graph where concepts are organized into a hierarchy in which more specific GO terms are subsumed by more general terms by following *is a* or in some cases *part of* relationships. The Evidence

Code Ontology (ECO) is a controlled vocabulary used for describing the scientific evidence that supports an assertion. ECO's 290+ terms include descriptions of laboratory experiments, computational methods and literature annotation terminology. The NCBI Taxonomy (TAXON) [9] is a database of taxonomic lineage obtained from a variety of sources, including primary literature, external databases and expert human curation efforts for databases hosted by the NCBI. The Sequence Ontology (SO) [12] describes a rich set of features and attributes of biological sequences. The terms and relations included in this ontology characterize both physical attributes of biological sequences (*i.e.* binding sites, exons) and the processes in which biological sequences may be involved in (*i.e.* translational frameshifts, transitions, deletions, *etc*). The Semanticscience Integrated Ontology (SIO) provides a basic set of types and relations for describing objects, processes and attributes of biological entities. Fig. 1 shows how these ontologies are used within the Gene dataset, or are linked to the Gene dataset by virtue of mappings to SIO.

Table 1. Summary metrics of ontologies that can be used to reason over the NCBI Gene dataset: the Evidence Code Ontology (ECO), the Gene Ontology (GO), the NCBI Taxonomy (NT), the SemanticScience Integrated Ontology (SIO) and the Sequence Ontology (SO)

Ontology	**Classes**	**Object properties**	**subClassOf axioms**	**subPropertyOf axioms**	**DL expressivity**
ECO	297	2	453	0	ALC
GO	34403	6	63375	0	ALE
TAXON	1018210	15	1018204	0	AL(D)
SIO	1385	201	1729	207	SRIQ(D)
SO	2151	74	2602	9	SHI

3 Reasoning Tasks

In this section, we describe reasoning tasks over the Gene-World knowledge base that can be used to benchmark the performance of an OWL reasoner or SPARQL query system. After loading all the triples for the NCBI Gene and HomoloGene RDF datasets, as well as all ontologies listed in Table 1, the first benchmark task for an OWL reasoner would be to check the consistency of the combined knowledge base. While each component is expected to not contain any unsatisfiable classes, mappings between SIO and Gene, SO and Gene or the disjoint class axioms for the NCBI taxonomy ontology may lead to class or property unsatisfiability.

Below, we present a set of DL and SPARQL-DL queries over the combined knowledge base that may not give the complete set of results without reasoning support for some portion of OWL2-DL (there are no nominals in the knowledge base). GitHub Gists of all queries are available at [13].

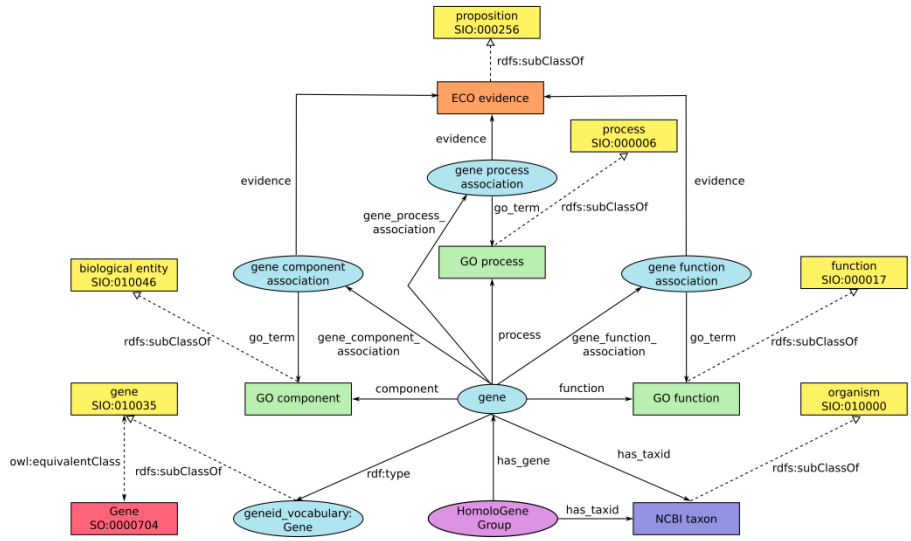

Fig. 1. Links between a gene in the NCBI Gene dataset and annotations of its function, associated cellular components and/or processes. Functions, cellular components, and processes are described using the Gene Ontology (GO, in green), while the associated evidence type for an association is described using the Evidence Codes Ontology (ECO, in orange). The taxonomic group for a gene is described using NCBI Taxonomy (in blue). HomoloGene (in purple) groups related genes and taxa. Each part of an NCBI Gene record is mapped to the Semanticscience Integrated Ontology (SIO, in yellow), which also has mappings to the Sequence Ontology (SO, in pink). Ellipses represent resources. Boxes represent ontology classes.

3.1 Query answering

Q1: retrieve transfer RNA genes
DL query: tRNA-gene
simple query that retrieves a type assertion in NCBI gene data

Q2: retrieve human genes
DL query: gene that has_taxid some ‘Homo sapiens [taxid:9606]’
conjunctive query over NCBI Gene and NCBI taxonomy

Q3: retrieve genes that are from any mammal but human
DL query: gene that has_taxid some (‘Mammalia [taxid: 40674]’ and not ‘Homo sapiens [taxid:9606]’)
conjunctive query with negation, and subclass reasoning over asserted hierarchy and class and relation mappings to upper level ontology

Q4: retrieve genes that are annotated with a specific enzymatic function:

DL query: gene that ‘has function’ some ‘acetylglucosaminyltransferase activity [go: 0008375]’
simple conjunctive query with subclass reasoning

Q5: retrieve genes that are annotated with a specific function that was not inferred by computational analysis.
DL query: gene that ‘has function’ some function that inverse(go_term) some (‘has evidence’ some (not ‘inferred from electronic annotation’))
conjunctive query using negation, mappings, inverse

Q6: retrieve organisms that have genes with an enzymatic activity that was not obtained by computational analysis
DL query: ‘Mammalia [taxid: 40674]’ that inverse(has_taxid) some (gene that 'has function' some (function that inverse(go_term) some ('has evidence' some (not 'inferred from electronic annotation')))
conjunctive query with negation, inverse, mappings

3.2 Querying using class axioms

All of the ontologies listed in Table 1 have rich class hierarchies. SIO and the Sequence Ontology (SO) also have axiomatic class definitions. DL queries can thus leverage the axioms used to define classes, as well as the class hierarchy.

Q7: retrieve a gene that encodes for a certain kind of molecule using SIO
DL query: gene and (encodes some ‘small cytoplasmic RNA (scRNA)’)
reasoning with subclass axioms from mapped ontology

Q8: retrieve a gene that encodes for a certain kind of molecule using SO
DL query: gene and (has_quality scRNA_encoding)
reasoning with subclass axioms from mapped ontology

3.3 SPARQL DL queries

SPARQL DL [14] is a subset of SPARQL that allows the formulation of queries using combination of OWL semantics and SPARQL variables. SPARQL DL is particularly useful in cases where one wishes to retrieve instances that are linked to some other resource, but also take advantage of DL reasoning. This is possible by using SPARQL variable bindings.

Q9: retrieve orthologous human and mouse genes annotated with function to bind ATP
Type(?human_gene, ‘gene’), Type(?mouse_gene, ‘gene’), Type(?homologene_group, HomoloGene_Group), PropertyValue(?human_gene, has_taxid, ‘Homo sapiens’), PropertyValue(?mouse_gene, has_taxid, ‘Mus musculus’), PropertyValue(?human_gene, ‘has function’, ‘ATP binding’),

PropertyValue(?mouse_gene, 'has function', 'ATP binding'), PropertyValue(?homologene_group, has_gene, ?human_gene), PropertyValue(?homologene_group, has_gene, ?mouse_gene)

4 Summary

We have described Gene-World, a large gene-centric knowledge base consisting of Bio2RDF datasets with over 395 million statements linked to five bio-ontologies with varying degrees of DL expressivity. The size and complexity of this dataset in addition to the provided DL and SPARQL-DL queries may provide a useful benchmark against which to evaluate OWL reasoner capability and efficiency for life science datasets. Should this preliminary knowledge base become useful in reasoner evaluation, we expect to extend it include more of the 20+ datasets and hundreds of ontologies in Bio2RDF.

5 References

1. Fernandez-Suarez XM, Galperin MY: **The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection**. *Nucleic Acids Res* 2013, **41**(Database issue):D1-7.
2. Chuang HY, Hofree M, Ideker T: **A decade of systems biology**. *Annu Rev Cell Dev Biol* 2010, **26**:721-744.
3. Callahan A, Cruz-Toledo J, Dumontier M: **Ontology-Based Querying with Bio2RDF's Linked Open Data**. *Journal of Biomedical Semantics* 2013, **4**(Supplement 1):S1.
4. Dentler K, Cornet R, ten Teije A, de Keizer N: **Comparison of reasoners for large ontologies in the OWL 2 EL profile**. *Semantic Web* 2011, **2**(2):71-87.
5. Guo Y, Pan Z, Heflin J: **LUBM: A benchmark for OWL knowledge base systems**. *Web Semantics: Science, Services and Agents on the World Wide Web* 2005, **3**(2-3):158-182.
6. Stearns MQ, Price C, Spackman KA, Wang AY: **SNOMED clinical terms: overview of the development process and project status**. *Proc AMIA Symp* 2001:662-666.
7. **Gene-World: A large-scale gene-centric semantic web knowledge base for molecular biology** [http://semanticscience.org/projects/gene-world]
8. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2011, **39**(Database issue):D52-57.
9. **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2013, **41**(Database issue):D8-D20.
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
11. **OWL Export of GO DATABASE DAILY TERMDB** [http://archive.geneontology.org/latest-termdb/go_daily-termdb.owl.gz]
12. Eilbeck K, Lewis SE: **Sequence ontology annotation guide**. *Comp Funct Genomics* 2004, **5**(8):642-647.
13. **Gene-World DL and SPARQL-DL Queries** [http://semanticscience.org/projects/gene-world/gene-world-query-gists.html]
14. Sirin E, Parsia B: **SPARQL-DL: SPARQL Query for OWL-DL**. In: *3rd OWL Experiences and Directions Workshop (OWLED-2007).* Innsbruck, Austria; 2007.