SEMANTIC QUERY OF A FUNGAL ENZYME KNOWLEDGEBASE IN THE OWL–DL ENVIRONMENT

Baker C.J.O, Shaban-Nejad A., Haarslev V.
Concordia University , Montreal, Quebec, Canada

Introduction

With the substantial increase in stored scientific data of various types, a major challenge of the post-genomic era is to access the knowledge stored in a myriad of complex databases and other resources across the web. Making these resources available in a more structured way and achieving simplified semantic access to units of intersecting information from different databases is the motivation of this study. To this end, the FungalWeb Ontology (FWO) written in the Ontology Web Language (OWL-DL), representing fungal taxonomy (NCBI / NEWT) and enzyme attributes (BRENDA) are mapped to establish a knowledgebase of use to enzyme application scientists working in the field of fungal genomics. Semantic query of the knowledgebase to identify instances of bio-scientific literature reporting industrially relevant enzymes produced by specific fungal taxonomic groups is described. Physio-chemical and catalytic properties of Laccase enzymes (EC-Number 1.10.3.2) in the context of the fungal host are investigated. Enzyme substrates are described in the context of the chemical dictionary of small molecular entities (ChEBI). The new Racer Query Language (nRQL) is used for defining instance retrieval queries using description logics.

Motivation

The key technical requirements for the development of the semantic web for genomics include the provision of formal ontologies associated with web sites, automated agent systems, text mining technologies, and relational data analysis. Together these components can deliver a robust integrated platform to provide genomics knowledge through semantic access. Semantic access to and retrieval of fungal and enzyme information - which are distributed to literatures, different databases and knowledge bases – through the Fungal Web Ontology (FWOnt) is motivation of this study.

The FWOnt is being used as a part of FungalWeb project. FungalWeb brings together Québec's existing expertise in ontology, machine learning and natural language processing to build a semantic web and intelligent system for fungal genomics.
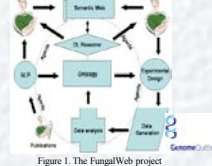
Figure 1. The FungalWeb project

Knowledge resources

The following resources along with additional distributed literature are the major data sources for FWOnt :

• **NCBI taxonomy database**: contains the names of all organisms including fungi that are represented in the genetic databases with at least one nucleotide or protein sequence.

• **NEWT** : the taxonomy database maintained by the Swiss-Prot. It is used in conjunction with the NCBI for extracting data for fungi concepts and instances

•**BRENDA**: a database of fungal enzymes and enzyme features. It gives a representative overview of enzyme characteristics, attributes, and properties

• **SwissProt**: a protein sequence database which provides a highly curated annotations, a minimal level of redundancy and high level of integration with other databases

• **ChEBI**: a dictionary of 'small molecular entities'. It encompasses an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents and/or children are specified.

FungalWeb (FWOnt) Ontology Structure

The FungalWeb ontology (FWOnt) aims to be an integrated, large-scale, formal bio-ontology for the fungal genomics domain to enable the study of fungal enzymes in OWL/DL environment. It reuses and integrates parts of other existing bio-ontologies such as GO and TAMBIS, and shares common concepts.
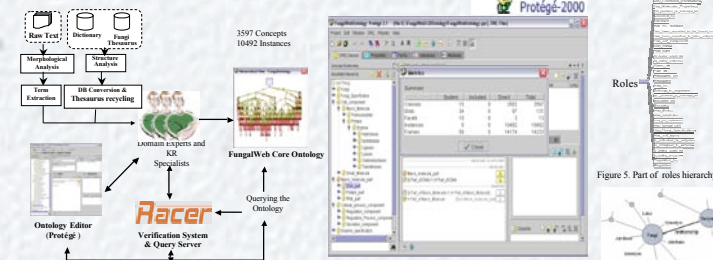
Figure 2. FungalWeb Ontology development life cycle

Figure 3. The FungalWeb Ontology in Protégé

Figure 4. An example of FWOnt Conceptual model

Figure 5. Part of roles hierarchy

Maximum expressiveness, without losing computational completeness and decidability of reasoning systems, is achieved using OWL-DL - a sublanguage of Ontology Web Language (OWL) with correspondence to description logics (DL). Additionally Protégé 2000 was used (with Owl plug-in) as our knowledge representation editor.

Example for concept description in DL: Basidiomaycota $\_$ Fungi $\sqsubseteq$ as_body.Fruity_body

The current taxonomy of the ontology is a full subsumption. The basic units in our taxonomy are instances of fungal species. By adding instances to our ontology, we create a knowledgebase instead of an ontology. How ever the line between ontology and knowledgebase is very narrow. Currently the FWOnt contains almost 3597 concepts and 10492 instances and still growing. All querying within ontology is done base on the FWOnt conceptualization (Figure 4, 5).

Semantic Querying

The Semantic Web is expected to include many kinds of query-answering services with access to numerous types of information represented in widely different formats.

Figure 6 shows the different biological data sources; NCBI, NEWT, ChEBI, SwissProt and BRENDA used in FungalWeb. To provide access to these integrated data, mapping of database entities to ontological concepts is necessary. In the domain of interest, these entities are enzymes, organisms (fungal species) and enzyme properties. Entity mapping to concepts in the ontology facilitates a variety of complex queries .
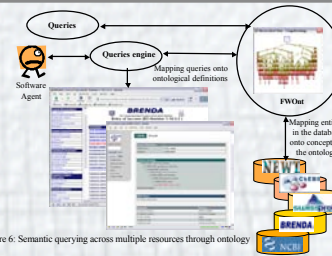
Figure 6: Semantic querying across multiple resources through ontology

Industrial Application Scenario for Semantic Querying

New environmental legislation forces a pulp manufacturer to consider reducing the chlorine used for delignification of pulp, implementing new bleaching technology or be fined:
• Alternatives to chlorine delignification include enzyme treatments.
• Laccase is known to bleach pulp when use in conjunction with mediator compounds.
• Pulp liquor is highly alkaline at the treatment step before chlorine bleaching normally occurs and enzyme treatment with laccase is only possible if the laccase can tolerate and operate in alkaline conditions.
• Since the majority of laccases have pH profile of 3-6 it would seem unlikely that laccase could help.
• How can the Project Manager find out if there is any report of a laccase with a pH optimum of 9.0
• If there is such an enzyme what species of organisms does this enzyme come from.
• How can I produce it in large quantities required for industrial scale bleaching
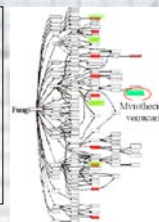
Figure 7: Visualizing the Query result for retrieving all Fungi which has been reported to have enzyme **Laccase** with **pH optimum of 9.0** and acts on the **Phenol**

Semantic Queries

The FungalWeb Ontology also provides the facility to connect with chemical databases like ChEBI. Figure 13, shows an enzyme substrate described in the context of the chemical dictionary of small molecular entities (ChEBI).

Figure 8: Communicating With ChEBI through enzyme substrate

RACER + nRQL

RACER is a DLs reasoning system with support for T-Box (concepts) and A-Box (instances)

• To check the scientific integrity of an ontology it in necessary to pose queries to the ontology.
• nRQL (New RACER Query Language) is a pragmatic A-Box Query language with formal semantics and inherits from the semantics of standard DL A-Box retrieval functions. nRQL unifies in a uniform declarative way the majority of RACER's ABox querying functions.
• Using RACER+ nRQL as a Semantic Web repository enables the constructuin of more flexible queries in domain of interest.

Figure 9, 10 demonstrate some sample queries and answers written in nRQL

Figure 9: nRQL Sample Queries

Figure 10: nRQL answers

Racer and nRQL documentation is available at: http://www.cse.concordia.ca/%7Ehaarslev/racer/download.html

Conclusion

• Semantic access and retrieval of heterogeneous data from distributed resources is a key focus of the Semantic Web

• Ontology drives Semantic Access by providing common conceptualization for use by people and machines

• Ongoing research involves improvement of querying capability through:

1. An Integrated global view and schema of all concepts / instances
2. User formulated query using ontology defined terms
3. System reformulation of queries to produce sub-queries for each resource
4. User gets answers with minimal knowledge of query methodology

Acknowledgments