
The FungalWeb Ontology: Application Scenarios

* Christopher J. O. Baker, René Witte, Arash Shaban-Nejad, Greg Butler and Volker Haarslev

Concordia University, Montreal, Quebec, Canada

ABSTRACT

Motivation: The FungalWeb Ontology aims to support the data integration needs of enzyme biotechnology from inception to product roll out. Serving as a knowledge base for decision support, the conceptualization seeks to link fungal species with enzymes, enzyme substrates, enzyme classifications, enzyme modifications, enzyme retail and applications. We demonstrate how the FungalWeb Ontology supports this remit by presenting application scenarios, conceptualizations of the ontological frame able to support these scenarios and semantic queries typical of a Biotech Manager. Queries to the knowledge base are answered with description logic (DL) and automated reasoning tools.

1 INTRODUCTION

Fungi are microorganisms well known for the range of novel enzymes they produce and enzymes of fungal origin are now used in industrial processes which amount to billions of dollars of revenue annually. The path to product development, namely gene discovery, enzyme characterization, mutational improvement and industrial application is long and fraught with numerous hurdles, both with respect to the domain knowledge and technical challenges. In an RnD environment many decisions are frequently made on incomplete knowledge. The current need is to have an integrated framework for discovery and decision support. This must integrate data from laboratory research, data accessed from distributed database, web and textual resources as well as the results of bioinformatics computation. To provide a reliable semantic resource in a contemporary RnD environment the scientific and technical span of ontology must encapsulate a more interdisciplinary range of concepts. The full range of conceptualizations required for commercial enzymologists includes taxonomy, gene discovery, protein family classification, enzyme characterization, enzyme improvement, enzyme production, enzyme substrates, enzyme performance benchmarking, and market niche. Inclusion of such concepts and instance data in ontology is within the scope of the FungalWeb data integration initiative.

2 ONTOLOGY DEVELOPMENT

The Fungal Web Ontology [1] is the result of integrating numerous biological database schema, web accessible textual resources and interviews with domain experts. The ontology includes both hierarchical structures supporting full-subsumption taxonomies and a broader conceptual frame

with novel relationships for specific domain knowledge. The major resources for fungal terminologies and concepts come from the following sources: NCBI taxonomy and literature databases [2], NEWT: is the taxonomy database [3], BRENDA enzyme database [4], *Saccharomyces* Genome Database [5], *Neurospora crassa* Genome Database [6], Commercial Enzyme Vendor Web Resources and the Enzyme Nomenclature Database [7].

The FungalWeb Ontology (FWOnt) reuses and integrates existing bio-ontologies and knowledgebases by merging, mapping and sharing common concepts using logics. Our ontology is an integrated ontology which used components of Gene ontology (GO) [8], TAMBIS [9] to establish the basic frame upon which biotechnology specific concepts have been added. The Ontology is a formal ontology written in OWL-DL, a sublanguage of Ontology Web language (OWL) with correspondence to description logics (DL). This provides maximum expressiveness, without losing computational completeness and decidability of reasoning systems. Protégé 2000 [10] was used (with Owl plug-in) as a knowledge representation editor. Aptness (considering completeness, consistency and conciseness) of the ontology for its intended application and the scientific integrity was evaluated by posing DL queries. RACER [11] was used as a description logic reasoning system with support for T-Box (concepts) and A-Box (instances).

3 APPLICATION SCENARIOS

We demonstrate the scope of our ontological conceptualization and the range of cross disciplinary queries that can be posed. We describe *junction* scenarios where a biotechnologist would ask support from the ontology and illustrate the how the diverse needs of the fungal biotechnology manager can be accommodated. The scientific context of these semantic queries and the conceptual frames designed to support them are outlined. nRQL syntax of DL queries to the ontology using Racer are presented for each scenario.

1.1 Enzymes acting on substrates

The ontology includes a concept representing the semantic stem of the systematic chemical names of enzyme substrates. This concept is instantiated with an NLP derived word stem of the most common term found in the enzyme descriptions of enzyme reaction classification scheme of the International Union of Biochemistry (IUB). By instantiating the semantically rich descriptions of the IUB into the conceptualization of the ontology we are able to query for mul-

tiple enzymes families known to degrade / modify a chemical substrate. A use case example querying for enzymes that act on the glucuronic acid polymer 'pectin' is described.

1.2 Enzyme provenance

A deep fungal taxonomy and enzyme reaction hierarchy are included in the ontology. The establishment of the relationship 'has been reported to be found in' between the concepts enzyme and fungus, reflecting information in scientific papers, facilitates the query of provenance of fungal enzymes (which enzyme is found in which fungal species). Such a query is further complemented by queries able to identify the common taxonomic lineage of all enzymes with a particular function and is of value to the biotech manager interested in the gene discovery and biodiversity.

1.3 Enzyme benchmark testing

An industrial specification is an important component of the FungalWeb ontology, representing concepts of value to the commercially oriented enzymologist. Access to information on commercial enzymes, product names, product parameters and vendors assists in the benchmarking the performance of newly discovered or mutationally improved enzymes. Typically such information is distributed on diversely formatted and company websites and promotional literature.

Fig. 1. Instance data generated by Mutation Miner.

```
<Protein>
  <Name>xylanase</Name>
  <Organisms>
    <Name>Bacillus circulans</Name>
  <label>PMID: 9930661: GI: 17942986</label>
  <Mark>D37N</Mark>
  <Context>The upward shift of the optimum pH of the D37N mutant was predictable from the results of structural and amino acid sequence comparison.
</Context>
```

1.4 Enzyme Improvement

An additional need of the commercial enzymologist is access to information on mutational studies resulting in better enzymes. We discuss the inclusion of ontological concepts to support the instance data produced by the NLP tools [12] designed specifically to extract information on experimentally introduced mutations and their impact on protein performance. The ultimate goal being to interrogate the ontology regarding mutations resulting in improved enzyme performance under defined environmental conditions. Instances generated by the NLP tool are shown in Figure 1.

CONCLUSION

We have used semantic web technology to create ontology and a large knowledgebase in the domain of fungal biotechnology and genomics from trusted biological sources to provide unified semantic access to heterogeneous resources. We have demonstrated the capacity of the ontological con-

ceptualization through a series of queries. Since our target audience is the decision making industry manager, not necessarily skilled in data mining technologies, we strive to facilitate answers without requiring advanced knowledge of query methodologies. We reason that sizeable time saving is made by and justifies the conceptual development of the ontology and its instantiation. Our semantic interrogations of the knowledge base provide us with further insight into structures of queries that the bio-scientific domain demands, thereby showing us the limits of the DL query technologies so that we can enhance the capabilities of Racer and nRQL.

ACKNOWLEDGEMENTS

This work was financed in part through the Genome Quebec project *Ontologies, the semantic web and intelligent systems for genomics* (V. Haarslev and G. Butler).

REFERENCES

- [1] Sheban-Nejad A., Baker C. J. O., Butler G. Haarslev V. (2004) *The FungalWeb Ontology: The core of a Semantic Web Application for Fungal Genomics, 1st Canadian Semantic Web Interest Group Meeting (SWIG'04) Montreal, Quebec, Canada*
- [2] Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA (2000). *Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2000 Jan 1;28(1):10-4*
- [3] Phan, I. Q. H., Pilbout S. F., Fleischmann W. and Bairoch A. (2003) *NEWT, a new taxonomy portal, Nucleic Acids Research, Vol. 31, No. 13 3822-3823*
- [4] Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. (2004) *BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. Jan 1;32(Database issue):D431-3.*
- [5] Saccharomyces Genome Database <http://www.yeastgenome.org/>
- [6] *Neurospora crassa* Database (<http://www.broad.mit.edu/annotation/fungi/neurospora/>)
- [7] Bairoch A. *The ENZYME database in 2000* (2000) *Nucleic Acids Res 28:304-305*
- [8] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) *Gene ontology: tool for the unification of biology. Nat Genet, 25(1):25-9*
- [9] Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. (1998) *TAMBIS--Transparent Access to Multiple Bioinformatics Information Sources. Proc Int Conf Intell Syst Mol Biol. 1998;6:25-34*
- [10] Noy N. F., Sintek M., Decker S., Crubezy M., Fergerson R. W., & Musen M. A.. (2000) *Creating Semantic Web Contents with Protege-2000 IEEE Intelligent Systems 16(2):60-71,*
- [11] Haarslev V, Möller R. (2001) *Description of the RACER System and its Applications. Proceedings of the International Workshop on Description Logics (DL-2001). Stanford, USA*
- [12] Witte R and Baker C.J.O. (2005) *Combining Biological Databases and Text Mining to support New Bioinformatics Applications. A.Montoyo et al. (Eds.) NLDB 2005, LNCS 3513,310-321.*