Extrapolating and Fusing Knowledge about Fatty Acid Biosynthesis in Arabidopsis Thaliana to Oil Biosynthesis in Brassica Napus

K. Allen (1), P. Gabel (2)

(1) Syngenta Biotechnology, 3054 Cornwallis Rd, Research Triangle Park, NC 27709
(2) Arity Corporation, 200 Friberg Pkwy, Westborough, Ma 01781
peter.gabel@Arity.com

We are interested in the application of semantic web technologies to specific questions in the life sciences. The specific study area is oil biosynthesis in Brassica napus, or oilseed rape, the third leading source of vegetable oil and the second leading source of protein meal according to the UN Food and Agricultural Organization. In principle, a great deal could potentially be learned about the relevant pathways in Brassica by comparison to the much better studied Arabidopsis thaliana, a close relative. Oil biosynthesis is enormously enhanced in Brassica, compared to Arabidopsis, and so comparison of the two genera has the potential to provide substantial insight into what gene families and pathways have been modified or amplified in Brassica to give rise to the observed levels of oil production (up to 50% of seed weight), as well as the differing oil composition.

An outline of our knowledge extrapolation and fusion process for this problem is the following:

a. Use GO, and GO Annotations rooted at "fatty acid biosynthesis" to identify gene products in Arabidopsis thaliana;
b. Use BLAST to identify gene products in Brassica napus that putatively correspond to selected Arabidopsis gene products
c. Use GO Annotations for Brassica napus
d. Identify and mine literature using gene names (There is literature available for genes and pathways in both genera), GO terms, and both species names. Synonyms will be integrated for gene/protein names and for GO terms.
e. Compute associations using ontological similarity kernels. This will provide clusters and confidence-based associations.

The Semantic Web technologies that are being used in this study are:

a) Inference of indirect relationships within the GO ontology using the OWL 1.1 (proposed) description logic fragment *EL++*. To the best of our knowledge we have the only implementation of full *EL++*.
b) Semantic indexing and retrieval of GO Annotations including clustering. Non-standard DL inferences and query processing of the data using our tools provide views of that are very suggestive of interesting structure in the data.
c) Structuring BLAST searches and results (attempting to use a *style* reflecting a Semantic Web approach)
d) Literature information extraction (while not strictly a Semantic Web technology, it reflects an important collateral technology)
e) Knowledge fusion that illustrates where SWRL and SPARQL may have value and, perhaps more importantly noted, where they do not apply. Information-theoretic ontology graph kernel functions provide similarity metrics which guide the knowledge fusion and matching process.

We are using this approach to strengthen ortholog assignments, compare gene families, discover relationships between gene products, and shed light on the gene families and pathways involved in oil biosynthesis.