Knowledge-Based Access to the Bio-Medical Literature

Ontologically-Grounded Experiments for the TREC 2003 Genomics Track

Richard Tong
Tarragon Consulting Corporation
1563 Solano Avenue, #350
Berkeley, CA 94707
rtong@tgncorp.com

John Quackenbush
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
johnq@tigr.org

Mark Snuffin
DataNaut, Inc.
11 Dudley Court, Suite 100
Bethesda, MD 20814
snuffin@datanaut.com

Abstract

The Tarragon Consulting team participated in the primary task of the TREC 2003 Genomics Track. We used a combination of knowledge-engineering and corpus analysis to construct semantic models of the interactions between genes/proteins and other biological entities in the organism, and then used automatic methods to convert these models into evidential queries that could be executed by the K2 search engine from Verity, Inc. The primary goal of our participation in the Genomics Track was to establish a performance baseline using ontologically-grounded techniques that are scalable and implementable using current commercial retrieval technology. The results from both our official submissions and subsequent experiments demonstrate that good performance can be achieved using our approach.

Overall Approach

In our approach we focused on "function" as opposed to the other aspects of the basic biology of the gene and its protein products. That is, we interpreted the primary task to be one that requires us to identify the ways in which the gene/protein is involved in the organism's behavior, as opposed to one that simply requires us to identify that some property of the gene/protein is being discussed.

The framework for constructing our semantic models is an ontology that makes a set of core distinctions between: (a) the gene/protein subsystem; (b) the organism; (c) the interactions of the gene/protein subsystem with the organism; and, (d) the documents that report on the biological entities and processes. (See Figure