Poster submission ISMB/ECCB 2007

Working title

My first BioAID: heuristic support for hypothesis construction from literature

Marco Roos, Sophia Katrenko, Willem van Hage, Edgar Meij, Frans Verster, Scott Marshall, Pieter Adriaans
Adaptive Information Disclosure, Faculty of Science, University of Amsterdam
e-mail: roos@science.uva.nl

Motivation

Constructing a new hypothesis is often the first step for a new cycle of experiments. A typical approach to harvesting biological literature is to scan the results of a PubMed query and read what we think is most relevant. In this scenario, we are limited by the selection of papers and, for future applications, we are limited by our capacity to recall the knowledge we have gained. As part of the development of a 'virtual laboratory for bioinformatics' (http://www.vl-e.nl), we seek alternative ways to support the construction of hypotheses from biological literature.

Objectives

Our objective is to provide automated support for hypothesis formation from literature based on an initial seed of knowledge.

Approach

Our approach consists of the following steps: first we create a 'proto-ontology' from the knowledge that we want to extend, for instance, a table in a review that lists diseases associated with a particular enzyme. We then identify the collection of documents that we want to search (typically Medline). Subsequently, we use concepts from our proto-ontology as input to retrieve relevant documents from a collection and to inform us of concepts such as protein names or relationships that are putatively associated with the proto-ontology. These results are used to enrich the proto-ontology with additional concepts and relations. The ontology can be iteratively enriched by using the results from one run as input for the next.

Implementation

Our implementation is based on a collection of web services, allowing us to construct custom workflows for specific tasks. Together, these web services form a toolbox called AIDA (Adaptive Information Disclosure Application), for annotating documents, searching documents, discovering knowledge from documents, and storing ontological data. AIDA uses open source software such as Lucene for document retrieval (http://lucene.apache.org), and Sesame for handling ontologies (http://www.openrdf.org). For the purposes of this implementation, we have also used Taverna to construct our workflows (http://taverna.sourceforge.net) and Protégé (http://protege.stanford.edu).

Results

We have created workflows from services in the AIDA toolbox, and applied them to extend a proto-ontology with knowledge extracted from literature. Technically, the most challenging workflow uses our own proto-ontology as input for machine learning services, after which biological concepts are discovered that are related to terms from our own ontology. As a proof of concept, we have (re)discovered diseases that are known to be related to EZH2, an enzyme associated with gene regulation via chromatin remodelling. A second workflow which discovers genomics concepts is used to identify proteins that might present a previously unreported link between two biological concepts, e.g. histones and transcription factors. The proto-ontology and enriched ontology are written in the Web Ontology Language OWL, and stored in Sesame via another service from the toolbox.

Availability

Services and workflows are available from http://ws.adaptivedisclosure.org/BioAIDdemo1. Ontologies are available from http://rdf.adaptivedisclosure.org/BioAIDdemo1.

Conclusion

Workflows constructed from the AIDA toolbox can be used as an aid in constructing hypotheses from literature. We show that we can automatically extend a proto-ontology with new hypothetical concepts and relationships that bridge across the boundaries of single papers or biological subdomains. Our approach can be customized to particular domains and vocabularies through the choice of ontology and literature corpora.