Statistics 290 Final Project: The RCyc Package

Tomer Altman

March 21, 2012

1 Introduction

Pathway Tools is a software platform for biochemical pathway bioinformatics [1]. It is closely tied to the EcoCyc and MetaCyc Pathway/Genome Databases (PGDBs), along with the BioCyc collection of PGDBs from over 1700 sequenced genomes. Pathway Tools currently has API support for Java and Perl. I added API support for interfacing Pathway Tools with R via an R package called 'taltman1.RCyc'.

2 Aims

I have created an API for interfacing Pathway Tools with R, termed RCyc. Analogous to the JavaCyc and PerlCyc APIs, the RCyc package allows an R process and a Pathway Tools process to communicate over a socket connection. Beyond merely managing the socket connection, the RCyc package parses a variant of XML used for encoding simple nested lists of strings and numbers from Pathway Tools (which is based on Common Lisp) into equivalent data structures in R. Furthermore, the Generic Frame Protocol API functions for frame representation systems are accessible via a generic R function for calling Pathway Tools code (including standard Common Lisp functions) along with other useful functions available in the Pathway Tools Lisp API.

A second package, Rcelot, was planned for but was beyond the scope of a one-quarter project. It would take the primitive frame manipulation

functions from RCyc and extend them to build S4 object classes and object instances in R. The aim is to mirror the instance and class structure as found in the Ocelot frame representation system in a PGDB as S4 classes and objects in R. This would allow R programmers to use the existing powerful object-oriented features of the language for analysis of PGDBs.

3 Sample use case

The Bioinformatics Research Group at SRI Interational has recently been approached by members of the BioConductor core team regarding providing our databases in a form that can be easily accessed by R, as they had been doing with the Kyoto Encyclopedia of Genes and Genomes. Especially in the wake of KEGG restricting access to its data, the BioConductor team needs a reliable biochemical pathway annotation resource.

With the RCyc and Rcelot packages, a user could extract all of the frames out of a PGDB and import them as objects into an R session. Example analyses that could then be performed would be gene set enrichment analysis, genome annotation training, and plotting of numerical features of genes (or other objects). Furthermore, all of the objects from a PGDB can be serialized after import into a flat-file that can then be distributed as a R object-oriented database representation of the PGDB, meaning that others could analyze the PGDB data contents without needing to be simultaneously connected to an instance of Pathway Tools. The standards of the BioConductor project for R-friendly dataset design will be consulted to ensure that standards are followed to ease adoption of these databases.

4 Implemented Functions

setUpPathwayToolsApiDaemon Starts up the Pathway Tools API Daemon on a socket.

callPToolsFn Allows one to call any Common Lisp function plus functions documented in the Pathway Tools Lisp API documentation.

shutDownPathwayToolsApiDaemon Cleanly shuts down the Pathway Tools API Daemon.

5 Future Work

It is my intention to get feedback on this package and prepare it for submission to BioCoonductor. Future extensions include automatically creating R versions of Pathway Tools functions, and in implementing RCelot for easy manipulation of PGDB data from within R.

References

[1] Karp, Altman, et al. **Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology**. *Brief Bioinform.*, vol. 11, no. 1, pp. 40-79, Jan 2010.