Databases and ontologies

Querying and computing with BioCyc databases

Markus Krummenacker¹, Suzanne Paley¹, Lukas Mueller², Thomas Yan³ and Peter D. Karp¹,*

¹SRI International, Bioinformatics Research Group, EK207 Menlo Park, CA 94025, USA, ²Cornell University, Department of Plant Breeding and Genetics, Emerson Hall Ithaca, NY 14853, USA and ³Carnegie Institution of Washington, Department of Plant Biology, Stanford, CA 94305, USA

Received on April 12, 2005; revised and accepted on June 14, 2005
Advance Access publication June 16, 2005

ABSTRACT

Summary: We describe multiple methods for accessing and querying the complex and integrated cellular data in the BioCyc family of databases: access through multiple file formats, access through Application Program Interfaces (APIs) for LISP, Perl and Java, and SQL access through the BioWarehouse relational database.

Availability: The Pathway Tools software and 20 BioCyc DBs in Tiers 1 and 2 are freely available to academic users; fees apply to some types of commercial use. For download instructions see http://BioCyc.org/download.shtml

Supplementary information: For more details on programmatic access to BioCyc DBs, see http://bioinformatics.ai.sri.com/ptools/ptools-resources.html
Contact: pkarp@ai.sri.com

1 INTRODUCTION

BioCyc (see http://BioCyc.org/) is a collection of 161 Pathway/Genome DataBases (PGDBs) that represent cellular networks and genome information in a structured manner, to allow powerful computational analysis and manipulation of data. The highly curated Tier 1 PGDBs at the core of BioCyc are the EcoCyc and MetaCyc DBs (Karp et al., 2002a). They contain many experimentally elucidated metabolic pathways from Escherichia coli and other organisms. BioCyc is viewed and edited through Pathway Tools (Karp et al., 2002a), a software environment we have developed to query, display and edit information about each pathway and its component reactions, compounds, enzymes, protein complexes, genes, operons and regulation at the substrate and transcriptional level. Additionally, the data objects support literature references, evidence codes and links to external databases. The BioCyc schema attempts to faithfully capture biological concepts and the cross-links among widely differing types of data. Tiers 2 and 3 were computationally predicted by Pathways Tools. Tier 2 has undergone moderate curation, whereas the 139 DBs in Tier 3 have undergone no curation (note also that Tier 3 PGDBs are not yet available for programmatic access, but we expect they will be soon).

This article describes multiple methods that are exposed for querying BioCyc DBs programmatically. The same access mechanisms are available for the many PGDBs now being created by Pathway Tools users outside SRI, such as by TAIR for Arabidopsis thaliana (Mueller et al., 2003), and by SGD for Saccharomyces cerevisiae. These query methods will simplify the investigation of global questions about cellular networks.

2 SCHEMA AND DATA FILES

BioCyc uses an object-oriented database called a Frame Representation System (FRS), the schema for which has been described previously (Karp, 2000); see also Appendix A of (Paley et al., 2005). In short, every biological object (such as a compound or gene) is stored in a frame bearing a unique ID. A frame has slots, in which attributes and connections to other frames can be stored as values. Slots can store single or multiple values, and individual values can be annotated with comments or literature references. The frames are organized in a class hierarchy.

Pathway Tools can export BioCyc PGDBs in several formats: (1) A column-delimited format and attribute-value format are described in detail online. (http://brg.ai.sri.com/ptools/flatfile-format.html) These formats are attractive for import into spreadsheets or relational DBs, or for parsing by Perl scripts. (2) BioPAX (http://www.biopax.org/) format, which is an OWL RDF/XML-based format for exchange of pathway data. (3) SBML (http://www.sbml.org/) format, which is an XML-based format for capturing models of biochemical reaction networks.

3 PROGRAMMATIC QUERIES

APIs in three languages provide direct, programmatic access to BioCyc DBs within Pathway Tools. The shared APIs are based upon the Generic Frame Protocol (GFP). The most commonly used GFP functions have been summarized (http://bioinformatics.ai.sri.com/ptools/gfp.html) and detailed documentation of GFP is available. (http://www.ai.sri.com/~gfp/spec/paper/paper.html) Additional useful functions (http://bioinformatics.ai.sri.com/ptools/ptools-fns.html) retrieve complex relationships in PGDBs. SQL querying is possible through the BioWarehouse.

Due to space limitations, only a simple example can be given below, which is transliterated to three languages: LISP, Perl and...
SQL. The example query finds all enzymes for which ATP is an inhibitor.

### 3.1 LISP

Common Lisp is the native programming language of Pathway Tools and thus provides the richest environment for queries. The API consists of the commonly used GPF functions plus the additional useful relations, as referred to above. Many Lisp query examples are available. (http://bioinformatics.ai.sri.com/ptools/examples.lisp)

```lisp
(defun atp-inhibits ()
  ;; We check every instance of the class
  (loop for x in (get-class-all-instances |Enzymatic-Reactions|)
    ;; We test for whether the INHIBITORS-ALL slot contains the compound frame ATP
    when (member-slot-value-p x INHIBITORS-ALL ATP)
    ;; Whenever the test is positive, we collect
    ;; the value of the slot ENZYME. The collected values are returned as a list,
    ;; once the loop terminates.
    collect (get-slot-value x 'ENZYME))
)
;; invoking the query:
(select-organism :org-id 'ECOLI)
(atp-inhibits)
```

### 3.2 PerlCyc

PerlCyc (http://www.arabidopsis.org/tools/aracyc/perlcycc/) is a Perl API that allows Perl programmers to query and update data within a running Pathway Tools server. The communication between Pathway Tools and Perl occurs through a UNIX socket, and so both programs need to be executed on the same machine.

```perl
use perlcyc;
my $cyc = perlcyc -> new("ECOLI");
my @enzrxns = $cyc -> get_class_all_instances (
  |Enzymatic-Reactions|)
## We check every instance of the class
foreach my $er (@enzrxns) {
  ## We test for whether the INHIBITORS-ALL
  ## slot contains the compound frame ATP
  my $bool = $cyc -> member_slot_value_p($er, "Inhibitors-All", "Atp");
  if ($bool) {
    ## Whenever the test is positive, we collect
    ## the value of the slot ENZYME. The results
    ## are printed in the terminal.
    my $enz = $cyc -> get_slot_value($er, "Enzyme");
    print STDOUT "$enz\n";
  }
}
```

### 3.3 JavaCyc

JavaCyc (http://www.arabidopsis.org/tools/aracyc/javacyc/) is a Java analog of PerlCyc. JavaCyc also communicates with Pathway Tools through a UNIX socket. The example query is available online (http://bioinformatics.ai.sri.com/ptools/example-javacyc.html).

### 3.4 SQL access via BioWarehouse

BioWarehouse is a DB integration project (http://bioinformatics.ai.sri.com/biowarehouse/) that allows multiple DBs including BioCyc, SWISS-PROT, Genbank, NCBI Taxonomy and KEGG to be loaded within a relational DBMS server. BioWarehouse supports SQL queries to BioCyc DBs, and allows cross-DB queries and validations to be performed. A detailed description of the BioWarehouse schema is beyond the scope of this Application Note.

```sql
select distinct DBID.xid
from DBID, Protein, EnzymaticReaction,
     EnzReactionInhibitorActivator, Chemical, DataSet
where DataSet.name=EcoCyc
  and DataSet.wid=EnzymaticReaction.datasetwid
  and EnzymaticReaction.proteinwid = Protein.wid
  and EnzymaticReaction.wid = EnzReactionInhibitorActivator.enzymaticreactionwid
  and EnzReactionInhibitorActivator.compoundwid=Chemical.wid
  and EnzReactionInhibitorActivator.inhibitoractivate=I
  and Chemical.name=ATP
  and DBID.otherwid = Protein.wid
```

### ACKNOWLEDGEMENTS

We thank Jeremy Zucker for the SBML exporter and Thomas J. Lee for his SQL example. This work was supported by grants GM70065 and GM75742 from the NIH National Institute of General Medical Sciences.

Conflict of Interest: Krummenacker and Karp declare that they receive royalties from SRI licensing of BioCyc and Pathway Tools, and Paley declares that she receives royalties from SRI licensing of Pathway Tools.

### REFERENCES