

PATRIC Standard Operating Procedures for Pathway Database Creation Using Pathway Tools Software

Author: Harsha K. Rajasimha
Version: 1.2, September 13, 2006
Supersedes version: 1.1

Purpose

This SOP describes the automated and manual procedures used for generating Pathway Genome Databases (PGDBs) (3) for the PATRIC bacterial genomes starting from their GenBank genomic sequence and PATRIC annotations. The document aims to familiarize personnel with the input, processing and output aspects of generating PGDBs using the SRI International's Pathway Tools software.

Background

A Pathway Genome Database (PGDB) is a database that integrates genomic data with detailed functional annotations of the genome, such as descriptions of metabolic and signaling pathways, and of the genetic network. The Pathway Tools software (4, 12) supports creation, editing, querying, visualization, and analysis of PGDBs. The software also allows users to publish a PGDB on the Web for querying by the scientific community.

The Pathway Tools software computes and predicts metabolic and signaling pathways by comparing the input genomic functional annotations with MetaCyc (1, 5, 9), the highly curated reference PGDB set. Since the MetaCyc collection mostly consists of microbial genomes, the performance of the procedure is more likely to be better for bacterial genomes compared to plant or mammalian genomes.

Application

CIG software developers and curators associated with Pathways curation. It also applies to a wider scientific community within VBI who may be interested in metabolic and signaling pathways prediction, curation, visualization and dissemination.

Participants

The tasks described herein are typically performed by a software engineer working with a curator. Ideally, at least one of the participants has undergone pathway tools training.

Inputs/References

FASTA sequence file for each replicon in the genome. A .pf (PathoLogic Formatted) annotation file each for all the replicons in the genome. A sample file in this format looks like this:

```
//  
ID      143026  
NAME    VBI0040CB1_1617  
STARTBASE 1491599  
ENDBASE  1492165  
FUNCTION hypothetical protein  
PRODUCT-TYPE P  
FUNCTION-COMMENT hypothetical protein  
GENE-COMMENT GB:NC_002971.3.472  
//
```

```

ID      141904
NAME    VBI0040CB1_1207
STARTBASE  1086288
ENDBASE  1086641
FUNCTION  hypothetical protein
PRODUCT-TYPE  P
FUNCTION-COMMENT  hypothetical protein
GENE-COMMENT  GB:NC_002971.3.472
//

```

Identify the ID for the PGDB being created. The typical naming format of a PGDB ID is *Cyc. E.g., EcoCyc (6, 8) for *E. coli* and AraCyc (14) for *Arabidopsis*. Also decide on the author list for PGDB and a footnote for all pathway pages.

Entry criteria

The pathway tools software is installed and configured for access to Oracle pathway curation database instance "BRCGUSCU" on the database server "Tuor". At least one of the participants should have experience with creating an Oracle PGDB. The file genetic-elements.dat file located at ptools-local/ directory (typically located on C:\ on Windows) should be edited to specify the correct .fasta and .pf file names identified in Section "Inputs/References". Ideally, one of the participants would have gone through the Pathway Tools tutorial/training. Otherwise, refer to <http://bioinformatics.ai.sri.com/ptools/> (4, 12).

Steps to create PGDB for a given genome starting from PATRIC annotations

1. Invoke the pathway-tools software by clicking the pathway-tools.bat file (typically on a curator's windows desktop).
2. From the menu bar, choose tools → PathoLogic. The PathoLogic module window opens up.
3. In the PathoLogic window that opens up, choose from the menu bar, Organism → Create New. A form opens up.
4. Fill in the form with all appropriate information you may have gathered in the "Inputs/References" phase and click OK. This will create a new PGDB directory with the specified name, typically in your ptools-local\pgdb\user\ directory.
5. Place all the input files (FASTA sequences .fsa or .fna and pathologic formatted annotation files .pf) appropriately named in the input directory of the PGDB thus created. Note that sequence file is optional to create a PGDB.
6. Edit the genetic-elements.dat file to reflect the correct input file(s). For example, the genetic-elements.dat file for *Coxiella burnetii* looks like this:

```

ID      GENOME40-CHROM-1
NAME    Coxiella Burnetii RSA 493 Chromosome 1
TYPE    :CHRSM
CIRCULAR?  Y
ANNOT-FILE  Coxiella_burnetii_RSA_493.NC_002971.pf
SEQ-FILE   493chr1.fna
//
ID      GENOME40-PLASMID-pQpH1
NAME    Coxiella Burnetii RSA 493 Plasmid pQpH1
TYPE    :PLASMID
CIRCULAR?  Y
ANNOT-FILE  Coxiella_burnetii_RSA_493.NC_004704.pf
SEQ-FILE   493chr2.fna

```

//

7. Now, you may test if the new PGDB has been created accurately as desired by choosing Build → Trial Parse from the menu on PathoLogic window. If all files are named and specified correctly, and if the annotation file in .pf format is syntactically error-free, you will see a message... “Trial Parse successful...” Otherwise, you may have to check the input files for syntactical correctness. For details on the pathway tools file formats, refer to <http://bioinformatics.ai.sri.com/ptools/>.
8. You may now proceed to Build → Automated Build step on the PathoLogic menu. This may take a few minutes and when complete, the automated PGDB file-KB is created in the PGDB directory at (typically at ptools-local\pgdbs\user\pgdbCyc). Note that during this step, the pathway tools software is actually comparing the function name of each feature with function names associated with known EC numbers in the reference PGDB set (MetaCyc). If the input annotation file(s) contain EC numbers to begin with, then these EC annotations are just directly used for constructing pathways.
9. The PGDB is still incomplete. To obtain a first version of the PGDB, you will need to perform all the tasks listed in the menu items under Refine → menu on the PathoLogic window. All steps up until here are typically performed by a software engineer or a bioinformatician.
10. In a sequential order, click each of the Refine → menu items, one at a time, and follow the instructions for manual validation (if any) at the end of each step. These tasks include:
 - a. Resolve Ambiguous Name Matches,
 - b. Assign Probably Enzymes,
 - c. Assign Modified Proteins,
 - d. Create Protein Complexes,
 - e. Re-Run Name Matcher,
 - f. Rescore Pathways,
 - g. Predict transcription units (7),
 - h. Transport Identification Parser,
 - i. Run Consistency Checker,
 - j. Update Overview (12), and
 - k. Pathway Hole Filler (2). It is only in this last step of identifying and filling holes with possible genome features, when sequence homology is used. This step requires that the software “blastall” be installed on your machine and included in the “PATH” environment variable.

Version 2.0 of this SOP should include more details on the manual intervention required in each of these steps.

All the above tasks are typically performed by a curator or the corresponding organism expert. At the end of this procedure, we should have the first version of PGDB for the organism in its directory within ptools-local. The PGDB thus created, is in its default format of ocelot-knowledgebase (KB). Ocelot is an SQL database management system. If the PGDB is intended for self use by the creator on his local machine, then the ocelot-KB format should suffice. However, for collaborative or concurrent editing (or curation) of these databases, the ocelot file-KB must be converted to either a MySQL or Oracle KB (13). For PATRIC (and all CIG projects), Oracle is the default RDBMS. Hence, you will need to choose from the PathoLogic organism menu, “Convert file-KB to Oracle-KB” item. Before doing so, make sure that the parameters are appropriately set in the ptools-init.dat file in your ptools-local directory. A sample entry is shown below:

```

# Parameters used by Pathway Tools when accessing PGDBs within a
# MySQL or Oracle server.
# -- Setting these parameters is MANDATORY if you want to access
# PGDBs from a MySQL or Oracle server.
# Hostname of the machine on which the MySQL or Oracle server is
# running.
RDBMS-Server-Hostname 128.173.97.5

# Port number on which either the MySQL or Oracle server is listening.
# MySQL standard is 3306; Oracle standard is 1521
#
RDBMS-Server-Port 521

# Name of database within the MySQL or Oracle server in which PGDBs
# should be stored.
#
RDBMS-Database-Name brcguscu.bioinformatics.vt.edu

# Username used to log in to the MySQL or Oracle server.
# If set to the value PROMPT, then the user will be prompted
# for their username and password.
#
RDBMS-Username myself

# Password used to log into the MySQL or Oracle server.
#
RDBMS-Password hi0015

```

Note that lines beginning with # are comments to user and will be ignored by the software. At this stage, you have a fully functional Oracle KB for the organism you started off with. You may now use the features of pathway tools software to browse, curate and visualize genomic, metabolic and signaling pathways for the organism.

One disadvantage of using the pathway tools software is the fact that we cannot query the PGDB (in any of the file, mysql, or oracle formats) using plain SQL queries. This is due to the implementation of the software in LISP, an artificial intelligence motivated language. However, application programming interfaces (APIs) have been developed in popular languages like Perl and Java for programmatic querying of PGDBs (10).

Measures

There is no definite metric for the performance of the procedure. The performance of the pathway tools software depends on the quality and accuracy of the underlying reference PGDB set (MetaCyc) and also on the quality of input genomic annotations. One measure of performance is the number of new EC annotations assigned by the procedure to previously hypothetical proteins. The authors of pathway tools software have shown an evaluation of the performance of computational prediction procedure for *Helicobacter pylori* (11).

Outputs

The output is a PGDB in two formats, namely, File-KB and Oracle-KB. The PGDB data can be exported into tab-delimited and attribute-value formats for distribution. Starting the pathway tools software in web-server mode makes all data visualization publicly available via a specified URL and port on the server.

Curators have a starting point of the PGDB for manual curation and validation.

Exit criteria

A single iteration through all steps outlined in the “Steps” section.

References

1. Caspi R, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Pick J, Rhee SY, Tissier C, Zhang P, Karp PD. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 1: D511-516, 2006.
2. Green ML. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5: 76, 2004.
3. Karp PD, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Kunin V, Lopez-Bigas N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 33: 6083-6089, 2005.
4. Karp PD, Romero P. The Pathway Tools software. *Bioinformatics* 18: s225-s232, 2002.
5. Karp PD, Paley SM, Pellegrini-Toole A. The MetaCyc Database. *Nucleic Acids Res* 30: 59-61, 2002.
6. Karp PD, Saier M, Paulsen IT, Collado-Vides J, Paley SM, and Pellegrini-Toole A, Gama-Castro S. The EcoCyc Database. *Nucleic Acids Res* 30: 56-58, 2002.
7. Karp PD. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway/genome databases. *Bioinformatics*, 2004.
8. Keseler IM, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Karp PD. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* 33: D334-D337, 2005.
9. Krieger CJ, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee and SY. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 32: D438-D442, 2004.
10. Krummenacker M, Mueller L, Yan T, Karp PD. Querying and computing with BioCyc databases. *Bioinformatics* 21: 3454-3455, 2005.
11. Paley SM. Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics* 18: 715-724, 2002.
12. Paley SM. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res* 34: 3779-3786, 2006.
13. Karp PD. What database management system(s) should be employed in bioinformatics applications? *OMICS* 7: 35-36, 2003.
14. Zhang P, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY. MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol* 138: 27-37, 2005.