How pathway databases were created and curated

Peifen Zhang
Plant Metabolic Network (PMN)
Introduction

The Plant Metabolic Network (PMN) is a collaborative project among databases and biochemists with a common goal to build a broad network of plant metabolic pathway databases. A central feature of the PMN is PlantCyc, a comprehensive plant biochemical pathway database, containing curated information from the literature and computational analyses about the genes, enzymes, compounds, reactions, and pathways involved in primary and secondary metabolism.

Item of the Month

The PMN launches PlantCyc 1.0

On June 17, 2008, PlantCyc made its web debut, containing biochemical pathways with information from over 290 plant species. The multi-organism database brings together over 2000
PMN is

• A network of plant metabolic pathway databases and database curation community
  – A plant reference database, PlantCyc
    • Genes, enzymes and pathways consolidated from all plant species

  – A collection of single-species pathway databases
    • Pathway Genome Databases (PGDB)
    • Genes, enzymes and pathways in a particular species

  – A community for data curation
    • Curators at databases (PMN, Gramene, SGN etc)
    • Researchers in the plant biochemistry field
Prediction of PGDBs, why

• Huge sequence data are generated from genome and EST projects

• Put individual genes into a metabolic network

• Use the network to visualize and analyze large experimental data sets, discover missing enzymes, design metabolic engineering, conduct comparative and evolutionary studies
Creation of PGDBs, how

- Manual extraction of pathways from the literature, assigning genes/enzymes to pathways

- Computational assigning genes/enzymes to reference pathways, manual validation/correction and further curation
Prediction of PGDBs, how

- Annotated sequences, molecular function
- A reference database (such as MetaCyc and PlantCyc)
- PathoLogic (Pathway Tools software)
PathoLogic

ANNOTATED GENOME

- DNA sequences
  - Gene calls
    - AT1G69370
  - Gene functions
    - chorismate mutase

- chorismate mutase
  - 5.4.99.5

- prephenate aminotransferase
  - 2.6.1.79

- arogenate dehydratase
  - 4.2.1.91

PGDB

- chorismate
- prephenate
- L-arogenate
- L-phenylalanine
A snapshot of AraCyc

• Arabidopsis genome
  – 27,235 protein coding genes

• AraCyc
  – 6158 enzyme coding genes
  – 2733 genes are assigned to reactions
  – 1914 genes are assigned to pathways
Currently available PGDBs

<table>
<thead>
<tr>
<th>Species</th>
<th>Database</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>TAIR</td>
<td>Substantial curation</td>
</tr>
<tr>
<td>Rice</td>
<td>Gramene</td>
<td>Some curation</td>
</tr>
<tr>
<td>Sorghum</td>
<td>Gramene</td>
<td>No curation</td>
</tr>
<tr>
<td>Medicago</td>
<td>Noble Foundation</td>
<td>some curation</td>
</tr>
<tr>
<td>Tomato</td>
<td>SGN</td>
<td>some curation</td>
</tr>
<tr>
<td>Potato</td>
<td>SGN</td>
<td>No curation</td>
</tr>
<tr>
<td>Pepper</td>
<td>SGN</td>
<td>No curation</td>
</tr>
<tr>
<td>Tobacco</td>
<td>SGN</td>
<td>No curation</td>
</tr>
<tr>
<td>Petunia</td>
<td>SGN</td>
<td>No curation</td>
</tr>
<tr>
<td>Coffee</td>
<td>SGN</td>
<td>No curation</td>
</tr>
</tbody>
</table>
Prediction of new PGDBs by PMN

• Prioritization
  – Available sequences, economic impact

• High priority
  – Maize, Poplar, Soybean, Wheat

• Second priority
  – Cotton, Grape, Sugarcane, Sunflower, Switchgrass…
A quality database REQUIRE manual validation and curation
Validation: pruning false-positive predictions

• Pathways not operating in plants or not in a target species
  – glycogen biosynthesis
  – C4 photosynthesis
  – caffeine biosynthesis

• Pathways operating via a different route
  – Phenylalanine biosynthesis in bacteria v.s. in plants
MetaCyc Pathway: phenylalanine biosynthesis

PlantCyc Pathway: phenylalanine biosynthesis
Validation: adding evidence and literature supports

*AraCyc Pathway: phenylalanine biosynthesis*

*AraCyc Pathway: ribose degradation*
Pathways are supported by different evidence

- Pathways supported by molecular data
  - enzymes and genes

- Pathways based on radio tracer experiments
  - no enzymes or genes

- Expert hypothesis (paper chemistry)

- Pure computational prediction
Correcting pathway diagrams
Curating missing pathways

• What information are curated from the literature
  – **Pathway**: diagram, summary, evidence, citations
  – **Reaction**: co-substrates, EC number
  – **Compound**: name and synonyms, structure
  – **Enzyme**: coding gene, physical-/biochemical properties, evidence, comments, citations
Source of literature

- PubMed, SciFinder
- Special journals (i.e. phytochemistry),
- Books in specialized field (i.e. alkaloids)
Curation workflow

1. Identify a pathway
2. Data entry
3. Find details of enzymes
4. Find details of reactions
5. Draw pathway diagram

- Reactions
- Species
- Structure of substrates
- EC number
- Enzymes
- Physical & chemical properties
- Coding gene
Current curation priority

- Big economic impact
  - Bio-energy production, i.e. cell wall components
  - Industrial material, i.e. rubber
  - Medicinal metabolites

- Under-represented domains
  - i.e. quinones, volatiles
The importance of community contribution, why we need your help

- A mountain of information
  - 17 million citations in PubMed alone
  - 4208 citations in PlantCyc

- Triage the most up-to-date and most relevant references

- Synthesize and extract information from individual papers
The importance of community contribution, why we need your help

• Limited human resource
  – curator (3 at PMN, 1 at SGN, 1 at Gramene)

• Limited expertise
  – molecular biologist, may be familiar in one particular pathway, but certainly not all the pathways.
How you can help

• Expedite data coverage
  – Submitting a pathway, an enzyme, a bunch of compounds

• Enhance data accuracy
  – Reporting errors

• Your idea/need of new features and functionalities
**Data submission forms**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Please SAVE this form and send as an ATTACHMENT to <a href="mailto:curator@plantcyc.org">curator@plantcyc.org</a></td>
<td></td>
<td></td>
<td></td>
<td>Pathway name (required)</td>
<td>Submission or Correction? (required)</td>
<td>(PATHWAY SUBMISSION / CORRECTION FORM)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pathway synonym(s) (required)</td>
<td>Organism(s) where the pathway exists (required)</td>
<td>Thank you for sharing your knowledge with us!</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Example: isoliquiritigenin biosynthesis</td>
<td>submission (new enzyme)</td>
<td>42′-trihydroxychalcone biosynthesis</td>
<td>Arabidopsis thaliana, medicago sativa, sesbania rostrata</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Example: isoliquiritigenin biosynthesis</td>
<td>submission (new enzyme)</td>
<td>same as row 4</td>
<td>same as row 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Please begin entering your data below:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
User Feedback Form

We welcome the comments and suggestions of our user community to help us maintain a high-quality and up-to-date resource. Please use the form below to report any of the following:

- An error or omission in the data
- An error or problem with a generated display page
- A suggestion for improvement
- Other comments or feedback

Alternatively, you may send email to curator@plantcyc.org.

Please fill in the following information:

Your Name: 
Your Email: 

URL where the problem appears:

http://www.plantcyc.org:1555/PLANT/NEW-IMAGE?type=REACTION&object=PREP-

Your comments, suggestions, or problem description:


Email to us

• curator@plantcyc.org
The PMN project, us and you

- **PlantCyc**
- **AraCyc**
- **MetaCyc**
- **tomato**
- **medicago**
- **rice**
- **poplar**
- **medicago**
- **sugarcane**
- **wheat**
- **maize**
- **other…**
Type of pathway databases

• Multi-species
  – MetaCyc (Universal, from microbes to plants to human)
  – PlantCyc (Plant kingdom)
  – BIACyc (a specific clade, for alkaloid biosynthesis)

• Single-species (Pathway Genome Database, PGDB)
  – AraCyc (Arabidopsis)
  – LycoCyc (tomato)
  – RiceCyc
  – etc