Top Four Essential TAIR Resources

Debbie Alexander

Metabolic Pathway Databases for Arabidopsis and Other Plants

Peifen Zhang
Metabolic Pathway Databases for Arabidopsis and Other Plants

Peifen Zhang
The Plant Metabolic Network (PMN)
Outline

• Introduction to PMN
• Query and retrieve the integrated information about pathways, metabolites, enzymes and genes
• Analyze sample gene expression data to identify changes in Arabidopsis metabolic pathways
• Behind the scene, how we create and curate a pathway database
http://plantcyc.org

encyclopedia
PMN is

- A network of plant metabolic pathway databases and database curation community

  - A plant reference database, PlantCyc
    - Pathways, enzymes and genes consolidated from all plant species

  - A collection of single-species pathway databases
    - Pathway Genome Databases (PGDB)
    - Pathways, enzymes and genes in a particular species

  - A community for data curation
    - Curators at databases (PMN, Gramene, SGN etc)
    - Researchers in the plant biochemistry field
Comparison between plantcyc and single-species PGDBs

• Plantcyc
  – Comprehensive collection of pathways for all plants
  – Representative collection of all known enzymes in plants

• A PGDB
  – Comprehensive collection of pathways in a particular species
  – Comprehensive collection of enzymes, known or predicted, in that species
The same underlying software, Pathway Tools

• Data creation, storage and exchange
• User interface
Major data types in PMN databases

- **Pathway**
  - Diagram, summary, evidence, species
- **Reaction**
  - Equation, EC number
- **Compound**
  - Synonyms, structure
- **Enzyme**
  - Assignments to reaction/pathway, evidence, summary, properties (inhibitor, Km etc)
- **Gene**
<table>
<thead>
<tr>
<th>PGDB</th>
<th>Species</th>
<th>Host</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraCyc</td>
<td>Arabidopsis</td>
<td>TAIR/PMN</td>
<td>Substantial curation</td>
</tr>
<tr>
<td>RiceCyc</td>
<td>Rice</td>
<td>Gramene</td>
<td>Some curation</td>
</tr>
<tr>
<td></td>
<td>Sorghum</td>
<td>Gramene</td>
<td>No curation</td>
</tr>
<tr>
<td>MedicCyc</td>
<td>Medicago</td>
<td>Noble Foundation</td>
<td>some curation</td>
</tr>
<tr>
<td>LycoCyc</td>
<td>Tomato</td>
<td>SGN</td>
<td>some curation</td>
</tr>
<tr>
<td></td>
<td>Potato</td>
<td>SGN</td>
<td>No curation</td>
</tr>
<tr>
<td></td>
<td>Pepper</td>
<td>SGN</td>
<td>No curation</td>
</tr>
<tr>
<td></td>
<td>Tobacco</td>
<td>SGN</td>
<td>No curation</td>
</tr>
<tr>
<td></td>
<td>Petunia</td>
<td>SGN</td>
<td>No curation</td>
</tr>
<tr>
<td></td>
<td>Coffee</td>
<td>SGN</td>
<td>No curation</td>
</tr>
</tbody>
</table>
Plant Metabolic Pathway Databases

**PlantCyc**

PlantCyc provides access to manually curated or reviewed information about shared and unique metabolic pathways present in over 250 plant species.

**AraCyc**

AraCyc provides access to manually curated or reviewed information about metabolic pathways for the model plant Arabidopsis thaliana. The pathways may be unique to Arabidopsis or shared with other organisms.

Data from gene expression, proteomic, and metabolomic experiments in Arabidopsis can be overlaid on a metabolic pathway map using the OMICS Viewer.

**External Plant Metabolic Databases**

Several additional species-specific databases, generated by PMN collaborators, are maintained and hosted at external sites. Some of the data from these databases have been incorporated into PlantCyc. Please click on the links below to access those databases directly, and read our Release Notes to learn about the content from external databases included in PlantCyc.

- CapCyc (Pepper) (Capsicum anuum)
- CoffeeCyc (Coffee) (Coffee canephora)
- LycosCyc (Tomato) (Solanum lycopersicum)
- MediciCyc (Medicago) (Medicago truncatula)
- NicotianaCyc (Tobacco) (Nicotiana tabacum)
- PetuniaCyc (Petunia) (Petunia hybrid)
- PotatoCyc (Potato) (Solanum tuberosum)
- RiceCyc (Rice) (Oryza sativa japonica Nipponbare)
- SolaCyc (Eggplant) (Solanum melongena)
- SorghumCyc (Sorghum) (Sorghum bicolor BTx623)
General query use cases

• How does a cell make or metabolize XXX?
• My gene is predicted to be a XXX, what does the enzyme do?
• What are the other genes involved in the same biochemical process as my gene?
• What is known and unknown of XXX pathway?
Pathway Tools Query Page

This form provides several different mechanisms for querying Pathway/Genome Databases.

Select a dataset: [PlantCyc] 3 available

- **Query**: All (by name or EC#)
  - To retrieve objects by name, first select the type of object you wish to retrieve, then enter the name of the object and click Submit. All objects containing the name as a substring will be returned. You may also enter multiple names or EC numbers, separating them with commas.

- **Browse Ontology**: [Pathways] Submit
  - Each dataset contains classification hierarchies for pathways, for reactions (the enzyme nomenclature system), for compounds, and for genes. Use this classification system to browse.

- **Choose from a list of all**: [Pathways] Submit

- **Cellular Overview**: [Diagram/Omics Viewer] (not available for MetaCyc)

- **Links to summary information about the selected organism**:
  - Summary page for dataset
  - History of updates to this dataset
  - Pathologic Pathway Analysis (not available for E. coli or MetaCyc)

- **Comparative Analysis**

caffeine
Pathways
Pathway pages contain: Depiction of metabolic pathway.

- caffeine biosynthesis I
- caffeine biosynthesis II (via paraxanthine)

Proteins
Protein pages contain: Detailed comments and citations; transcription factors.

- caffeine synthase (polypeptide)
- caffeine synthase (polypeptide) - CaDXMT1
- caffeine synthase (polypeptide) - CCS1
- caffeine synthase (polypeptide) - TCS1

Compounds
Compound pages contain: Compound structural inform

- caffeine

Synonyms: 1,3,7-trimethylxanthine
Empirical Formula: C_{8}H_{10}N_{4}O_{2}
Molecular Weight: 194.19 daltons

Smiles: c12(n(C)c(=O)n(C)c(=O)c(n(C)cn1)2)

In Pathway Reactions as a Product:
- caffeine biosynthesis II (via paraxanthine):
  - paraxanthine + S-adenosyl-L-methionine = caffeine + S-adenosyl-L-homocysteine
- caffeine biosynthesis I:
  - theobromine + S-adenosyl-L-methionine = caffeine + S-adenosyl-L-homocysteine
Caffeine, 1,3,7-trimethylxanthine, is one of the best known purine alkaloids and also the most abundant. Caffeine presents in high concentrations in many young expanding leaves of Coffea arabica, and up to 3% dry weight in young leaves of tea (Camellia spp.) plants is undetermined. Their hypothetic roles include being a defense chemical where high concentrations protect these soft tissues from predators, or being an autotoxic chemical where caffeine reoccurrence is little evidence supports the hypothesis.

The depicted pathway herein illustrates the major routes of caffeine biosynthesis, from xanthosine to 7-methylxanthine. Hypothetical enzymes specific xanthosine N-methyltransferase, 7-methylxanthine N-methyltransferase, and theobromine N-methyltransferase activity. It specifically converts xanthosine to 7-methylxanthine. Theobromine synthase has also been characterized from coffee which has the specific 7-methyltransferase activity (the second methylation step). Caffeine synthase has been characterized from both coffee and tea, N-methyltransferase and theobromine N-methyltransferase activities. It converts 7-methylxanthine to caffeine and theobromine. Immature fruits are the major sites of caffeine biosynthesis in coffee plants. Theobromine and paraxanthine activities are also found.

Free purine nucleotides are the major resources of xanthosine. Xanthosine is derived from nucleosides). Newly synthesized IMP from de novo purine biosynthesis, and adenosine reactivated are converted to xanthosine and enter the caffeine biosynthesis pathway.

In addition to the major route, caffeine may also be synthesized via a few minor routes such as paraxanthine.

Citations: [Misako04, Ogawa01, Mizuno03b, Mizuno03a, Kato96]
**PlantCyc Enzyme: chorismate mutase / L-ascorbate peroxidase**

**Enzymatic reaction of: chorismate mutase**

chorismate $\leftrightarrow$ prephenate

The reaction direction shown, that is, $A + B \leftrightarrow C + D$ versus $C + D \leftrightarrow A + B$, is unspecified.

Reversibility of this reaction is unspecified.

In Pathways: *tyrosine biosynthesis II*, *phenylalanine biosynthesis*, *tyrosine biosynthesis*

Activators (Allosteric): L-tryptophan [Mobley99], L-tryptophan [Mobley99]

Inhibitors (Allosteric): L-tryrosine [Mobley99], L-phenylalanine [Mobley99], L-phenylalanine [Mobley99]

Primary Physiological Regulators of Enzyme Activity: L-tryptophan, L-phenylalanine

$K_M$ for chorismate: 2900 $\mu$M [Mobley99]

**Enzymatic reaction of: L-ascorbate peroxidase**

$2 \text{ L-ascorbate} + \text{H}_2\text{O}_2 \leftrightarrow 2 \text{ monodehydroascorbate} + 2 \text{ H}_2\text{O}$

The reaction direction shown, that is, $A + B \leftrightarrow C + D$ versus $C + D \leftrightarrow A + B$, is unspecified.

Reversibility of this reaction is unspecified.

In Pathways: *ascorbate glutathione cycle*

References:

Outline

• Introduction to PMN
• Query and retrieve the integrated information about pathways, metabolites, enzymes and genes
• Analyze sample gene expression data to identify changes in Arabidopsis metabolic pathways
• Behind the scene, how we create and curate a pathway database
Omics Viewer

Overview of the AraCyc Metabolic Map
Omics Viewer use cases

• Visualize and interpret large scale omics data in a metabolism context
  – Gene expression data
  – Proteomic data
  – Metabolic profiling data
  – Reaction flux data
Step one: prepare data input file

Select a dataset: Arabidopsis thaliana col

File containing experimental data (NOT a URL):

Do you want to display absolute or relative data values?

If displaying relative data values, use:
- a single data column
- the ratio of two data columns
- 0-centered scale (e.g. log scale)
- 1-centered scale (negative values)

Data values use a:
- Gene names and/or identifiers

Note: By selecting Any of the above, you can combine, for example, metabolomics data into a single display. There are some dangers however. Some names may be ambiguous if it is not known if the metabolites. In addition, data values from different kinds of experiments may be comparable, so the resulting diagram may be misleading in some cases.
Step two: choose which data to be displayed

Single Experiment Time Step or Animated Time Series

To display a single experiment time step, enter a single column number in one or both of the column number fields below.

To display an animated time series, enter a list of column numbers (with each column number corresponding to a single timepoint), one per line, in the first column number field below. If you wish to include a denominator column for a ratio calculation, you can enter either a single column number (in which case the same data column will be used as the denominator for all timepoints), or one column number for each numerator column number. Note that zoomed views of individual pathways are not available with animations.

Data column (numerator in ratios): If using two columns, denominator data column:

Note: For column numbering purposes, the first column, which contains the gene name, is column number 0. The first potential data column is column number 1.
Last, choose color scheme and display type

Choose a color scheme:
- Full color spectrum, computed from data provided (default)
- Full color spectrum with a maximum cutoff: [ ]
- Three color display with specified threshold: [1]

**Display Type**

By default, data values are painted on the cellular overview chart. However, an alternative display is to either paint data values on the genomic map, or to generate a table containing all individual pathways which have one or more data values that exceed some threshold (or are less than the inverse of that threshold). To select one of these alternative displays, choose the corresponding option below and specify the threshold if appropriate. Note that if both the cellular and genome overviews are specified, the genome overview will appear in a new browser window (you must have popups enabled for this site or this will not work).
- Paint data on cellular overview chart (default)
- Paint data on genome overview chart
- Generate a table of individual pathways exceeding threshold: [ ]

**Submit**

Note that this request will take several minutes to complete (possibly longer for large datasets). For faster operation, install Pathway Tools on your own computer! [Click here](#) for details.
Red: enhanced expression over my threshold
Yellow: repressed expression over my threshold
Blue: not significant under my threshold
Close-up to a pathway of interest
Generate a table of individual pathways exceeding certain threshold.
Outline

• Introduction to PMN
• Query and retrieve the integrated information about pathways, metabolites, enzymes and genes
• Analyze sample gene expression data to identify changes in Arabidopsis metabolic pathways
• Behind the scene, how we create and curate a PGDB
Create PGDBs, why

• Huge sequence data are generated from genome and EST projects

• Put individual genes into the context of metabolic network

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

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

1. **Gene calls**
   - AT1G69370

2. **Gene functions**
   - chorismate mutase

**PathoLogic**

- chorismate mutase: 5.4.99.5
- prephenate aminotransferase: 2.6.1.79
- arogenate dehydratase: 4.2.1.91

**PGDB**

- chorismate
- prephenate: 2.6.1.79
- L-arogenate: 4.2.1.91
- L-phenylalanine

**MetaCyc**
New PGDB pipeline in PMN

• Prioritization
  – Available sequences, economic impact

• High priority
  – Poplar, Soybean, Maize, Wheat

• Others
  – Cotton, Grape, Sugarcane, Sunflower, Switchgrass…
A quality database requires manual validation and curation
Validation: prune 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
Validation: add evidence and literature support

- Molecular data, enzymes and genes
- Radio tracer experiments
- Expert hypothesis (paper chemistry)
- Pure computational prediction

*Aracyc Pathway: phenylalanine biosynthesis*
Curation: correct pathway diagrams

L-tryptophan → CYP79B2 tryptophan monoxygenase: At4g39950
O₂, NADPH
H₂O, NADP⁺, CO₂

indole-3-acetaldoxime → CYP83B1 monoxygenase: AT4G31500
L-cysteine, O₂, NADPH
NH₃, pyruvate, H₂O, NADP⁺

indolylmethylthiohydroximate → UDP-D-glucose, UDP

indolylmethyl-desulfoglucosinolate → PAPS, 2.8.2.

indolylmethyl-glucosinolate
Curation: correct gene/enzyme assignments to reaction/pathway

UGT89C1

Enzymatic reaction on: *flavonol 7-O-rhamnosyltransferase*

quercetin-3-glucoside + UDP-L-rhamnose <=> quercetin-3-O-glucoside-7-O-rhamnoside + UDP

The reaction direction shown, that is, \( A + B \rightleftharpoons C + D \) versus \( C + D \rightleftharpoons A + B \), is in accordance with the direction within a pathway.

Reversibility of this reaction is unspecified.

In Pathways: quercetin glucoside biosynthesis

Citations: [YonekuraSa07]

---

Enzymatic reaction on: *UDP-glycosyltransferase (flavonol 7-O-rhamnosyltransferase)*

trans-zeatin + UDP-D-glucose <=> trans-zeatin-7-N-glucoside + UDP

Reversibility of this reaction is unspecified.

In Pathways: cytokinins 7-N-glucoside biosynthesis
Further curation

• Add missing or new pathways
• Add missing or new enzymes
• Add detailed literature information about a pathway, an enzyme etc
Community curation

• Adopt a newly created PGDB by a genome database

• Participate as a lab/group

• Participate as an individual

• Contact us: curator@plantcyc.org
Thank you!