# TAIR/Gramene/SGN Workshop I

**ASPB Meeting
July 08, 2007
Chicago, IL**

## Metabolic Databases

# MetaCyc and AraCyc: Curation of Plant Metabolism

Hartmut Foerster

Carnegie Institution

Hartmut Foerster

Carnegie Institution

# Outline

- **MetaCyc**
  - Goals and application
  - Curation Progress
  - MetaCyc – main functions
  - Pathway tools

- **AraCyc**
  - Build of AraCyc
  - Curation progress
  - Introduction to the database
  - Omics viewer

# MetaCyc

http://www.metacyc.org

Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, Tissier CP, Zhang P, Karp PD
**MetaCyc: a multiorganism database of metabolic pathways and enzymes**
Nucleid Acids Res., 34, D511- 516  (2006)

# What is MetaCyc ?

- MetaCyc is a multi-organism database that collects any known pathways across all kingdoms

- MetaCyc is a curated, literature-based biochemical pathways database

- Collaboration between *SRI International* and *Carnegie Institution*

# Goal and Applications

## Goal

- Universal repository of metabolic pathways
  - Up-to-date, literature-curated catalogue of commented enzymes and pathways for use in research, metabolic engineering and education

## Applications

- Database of reference used to generate predicted Pathway/Genome DataBases (PGDBs)

# The Content of MetaCyc

- **Pathways** from primary and secondary (specialized) metabolism
  - Reactions with compound structures
    - Proteins
    - Genes

  - Does not contain sequence information

# Curation Team

- PhD-level curators

- Extract data from literature
  - experimentally verified data
  - ideally protein information

- Follow MetaCyc Curator's guide
  - *http://bioinformatics.ai.sri.com/ptools/curatorsguide.pdf*

# MetaCyc Version 11.1
## released May 25th, 2007



Note: The statistics for each year pertain to the last MetaCyc version released in that year

# The Taxonomic Distribution in MetaCyc

- Mostly from microorganism and plant kingdoms, and several animal pathways

- Plants: 219 species from 69 plant families

- Share about 220 pathways involved in primary metabolism and 180 pathways of secondary (specialized) metabolism
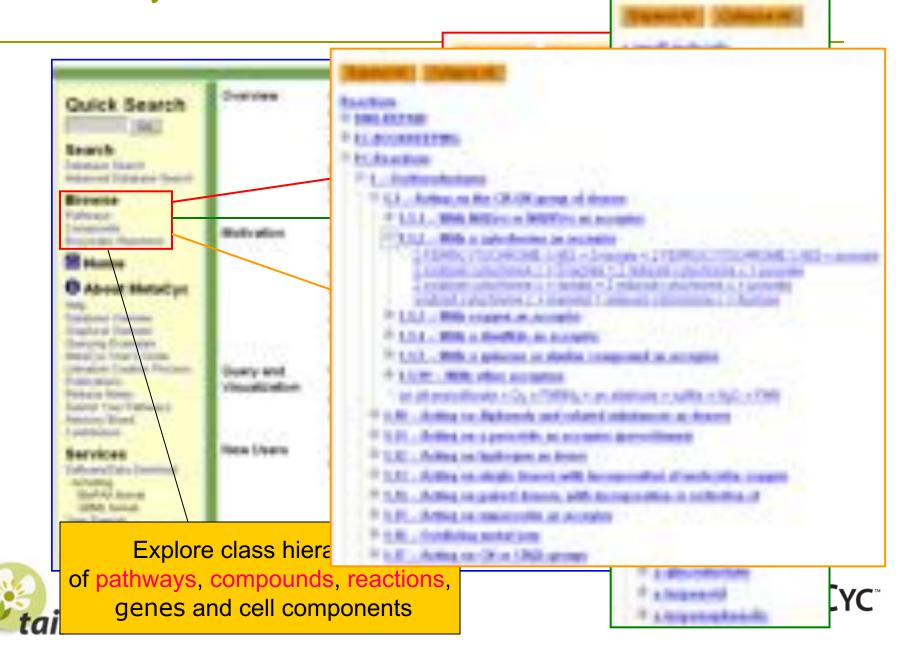
# Taxonomy/Metabolism Ratio in MetaCyc

- Plant families with the highest number of annotated pathways:

  Brassicaceae (*Arabidopsis thaliana)*

  Legumes        (*Glycine max*)

  Poaceae        (*Zea mays*)

  Solanaceae   (*Solanum tuberosum, Nicotiana tabacum)*

- Plant families with the highest number of contributing species:

  Legumes  (36)                    Solanaceae   (16)
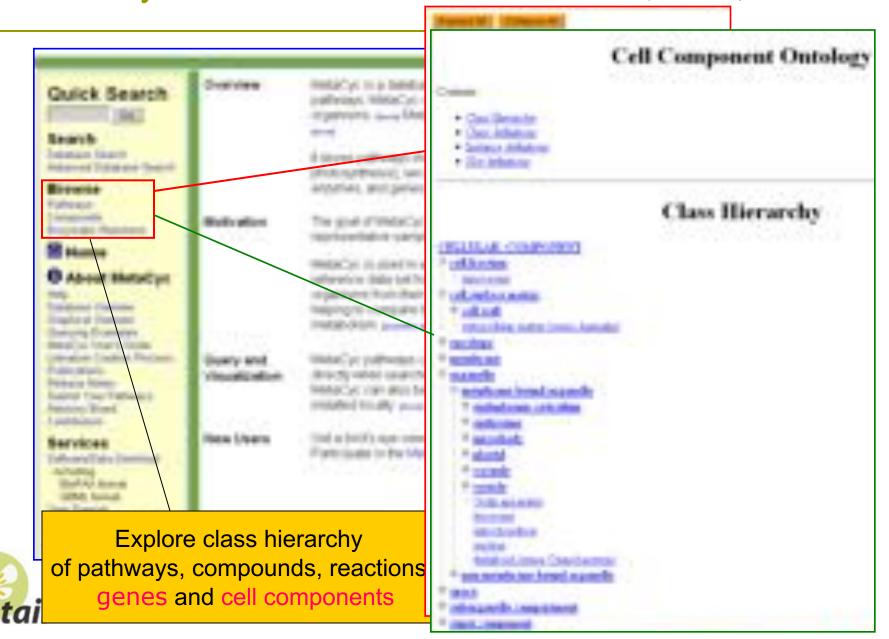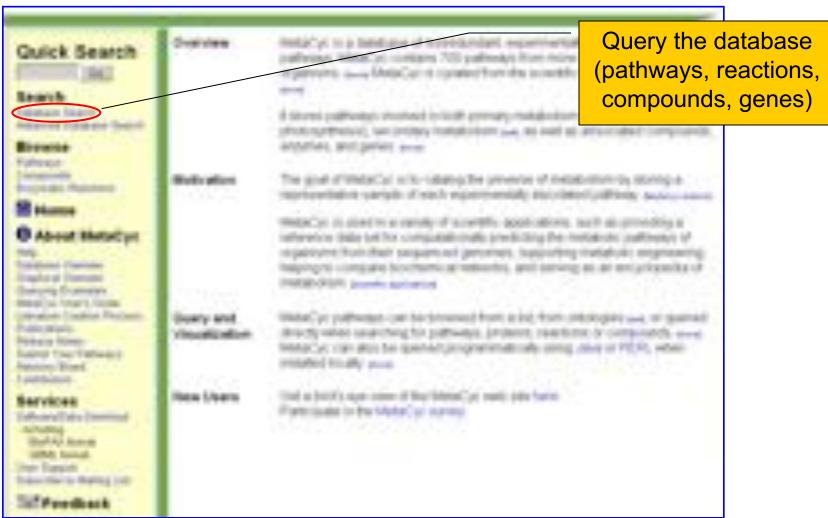  Poaceae   (12)                    Brassicaceae (10)

# MetaCyc – Browse the Database



Explore class hierarchy of pathways, compounds, reactions, genes and cell components

# MetaCyc – Browse the Database *(cont'd)*



Explore class hierarchy
of pathways, compounds, reactions,
genes and cell components

# MetaCyc – Discover the Metabolic Universe



Query the database (pathways, reactions, compounds, genes)

# MetaCyc – Query Page



Type your search term and click submit

# MetaCyc – Query Result Page

# MetaCyc – Pathway Detail Page - Part I



The pathway diagram shows **compounds**, **reactions** and **metabolic links**

# MetaCyc – Pathway Detail Page - Part II

Pathway commentary comprises general and specific information about the pathway

# MetaCyc – More Detail



Extend or collapse the detail level of the pathway detail page

# MetaCyc – More Detail *(cont'd)*



In depth information about reaction, EC number, enzymes, genes, regulatory aspects, and metabolic links to related pathways

# MetaCyc – More Detail *(cont'd)*

# MetaCyc – More Detail *(cont'd)*



More detail reveals structural information about compounds

# MetaCyc – Reaction Detail Page

# MetaCyc – Enzymes and Genes



Contains enzyme commentary, references, and physico-chemical properties of the enzyme

# MetaCyc – Enzymes and Genes *(cont'd)*

# Variants, Related Pathways and Links

- **Pathway variants are created as separate pathways**
  - *IAA biosynthesis I (tryptophane-dependent)*
  - *IAA biosynthesis II (tryptophane-independent)*

- **Links are added between interconnected pathways**
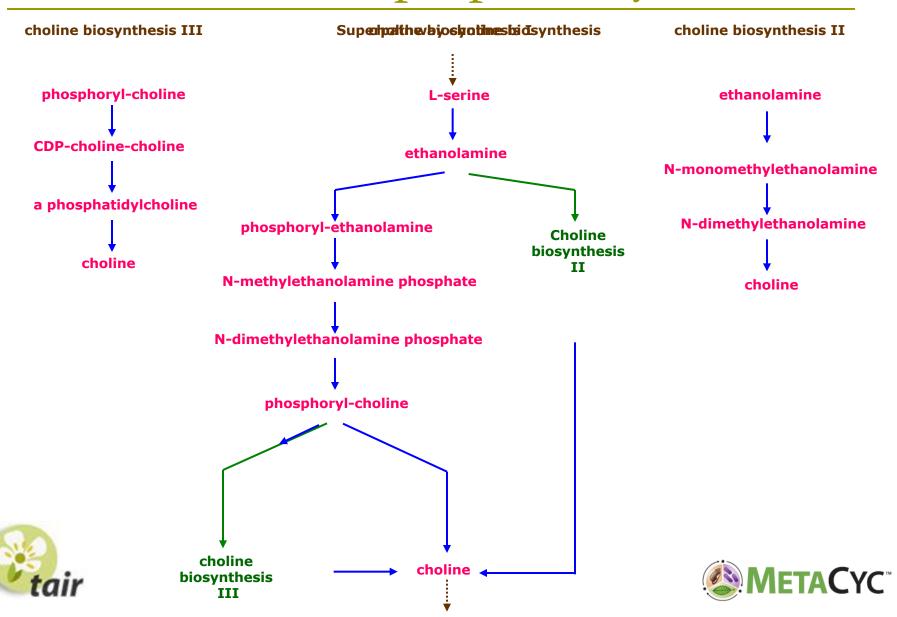
- **Related pathways are grouped into superpathways**
  - *e.g.superpathway of choline biosynthesis*

# Creation of a Superpathway

**choline biosynthesis III**

**Superpathway of choline biosynthesis**

**choline biosynthesis II**

**phosphoryl-choline**

↓

**CDP-choline-choline**

↓

**a phosphatidylcholine**

↓

**choline**

**L-serine**

↓

**ethanolamine**

**phosphoryl-ethanolamine**

↓

**N-methylethanolamine phosphate**

↓

**N-dimethylethanolamine phosphate**

↓

**phosphoryl-choline**

**Choline biosynthesis II**

**choline biosynthesis III**

**choline**

**ethanolamine**

↓

**N-monomethylethanolamine**

↓

**N-dimethylethanolamine**

↓

**choline**

# Applications: Pathways Prediction

## Goal

- Universal repository of metabolic pathways
  - Up-to-date, literature-curated catalogue of commented enzymes and pathways for use in research, metabolic engineering and education

## Applications

- Database of reference used to generate predicted Pathway/Genome DataBases (PGDBs)

# Pathway Tools Software Suite

- Software for generating, curating, querying, displaying PGDBs
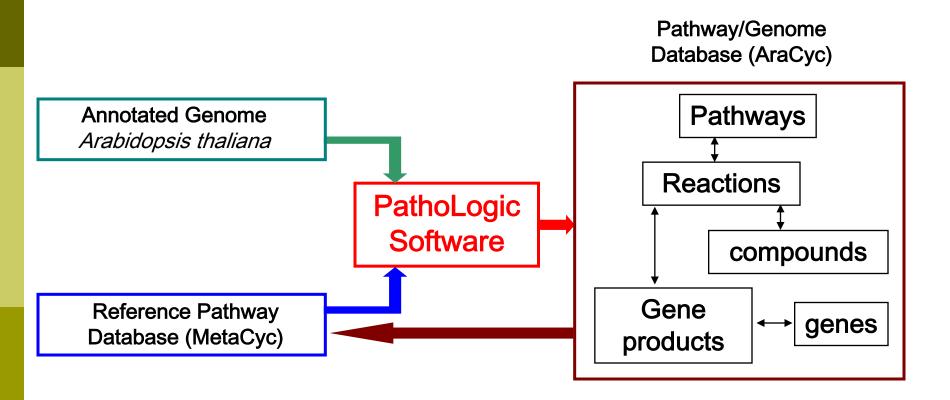
- Developed by Peter Karp and team
  - PathoLogic – Infers pathways from genome or transcripts sequencing
  - Pathway/Genome Editors – Curation interface
  - Pathway/Genome Navigator – Query, visualization, analysis and Web publishing
  - OMICS Viewer

# The Family of Species-specific Databases

Pathway/Genome
Database (AraCyc)

Annotated Genome
*Arabidopsis thaliana*

PathoLogic
Software

Reference Pathway
Database (MetaCyc)

Pathways

Reactions

compounds

Gene
products

genes

# AraCyc:
## The *Arabidopsis thaliana* specific metabolic database

http://www.arabidopsis.org/tools/aracyc

# AraCyc – Birth of the *A. thaliana* Specific Database
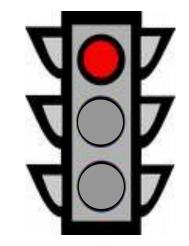
# The Computational Build of AraCyc

- In 2004, the *Arabidopsis* genome contained 7900 genes annotated to the GO term 'catalytic activity'

- 4900 loci in small molecule metabolism (19% of the total genome)

- PathoLogic inferred 219 pathways and mapped 940 (19% enzyme-coding) genes to the pathways

# Cleaning of a Newborn Database

- PathoLogic errs on the side of over-prediction

- First round of curation to remove false-positives

- Add missing pathways

- Improve the quality of information
    - Introduce new pathways
    - Increase number of pathway and protein comments
    - Refine computational assignment of protein

# Pathway Validation Criteria

- A pathway that is described in the *Arabidopsis* literature

- A pathway whose crucial metabolites are described in the *Arabidopsis* literature

- A pathway that contains unique reactions and having genes assigned to those unique reactions
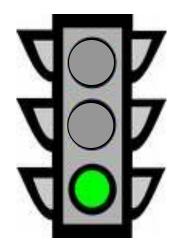
# Validation Procedure

- **Delete non-plant pathways:**
  - Pathway variants of bacteria-origin
  - Pathways not operating in plants at all (e.g. glycogen biosynthesis)

- **Add new plant-specific pathways:**
  - Pathway variants of plant-origin
  - Plant-specific metabolites
    (e.g. plant hormones)
  - Plant-specific metabolism
    (e.g. xanthophyll cycle)

# AraCyc - Curation Progress

| | AraCyc (2.1) April 2005 | AraCyc (2.5) October 2005 | AraCyc (2.6) May 2006 | AraCyc (3.5) February 2007 | AraCyc (4.0) July 2007 |
|---|---|---|---|---|---|
| Total pathways | 221 | 197 | 228 | 262 | 285 |
| New | - | 37 | 35 | 51 | 50 |
| Updated | - | 0 | 4 | 37 | 42 |
| Deleted | - | 61 | 6 | 12 | 16 |
| Pathways manually reviewed | 71 (32%) | 170 (86%) | 201 (88%) | 233 (89%) | 285 (100%) |

# How to Link to AraCyc



From the TAIR home page click on the link to AraCyc pathways
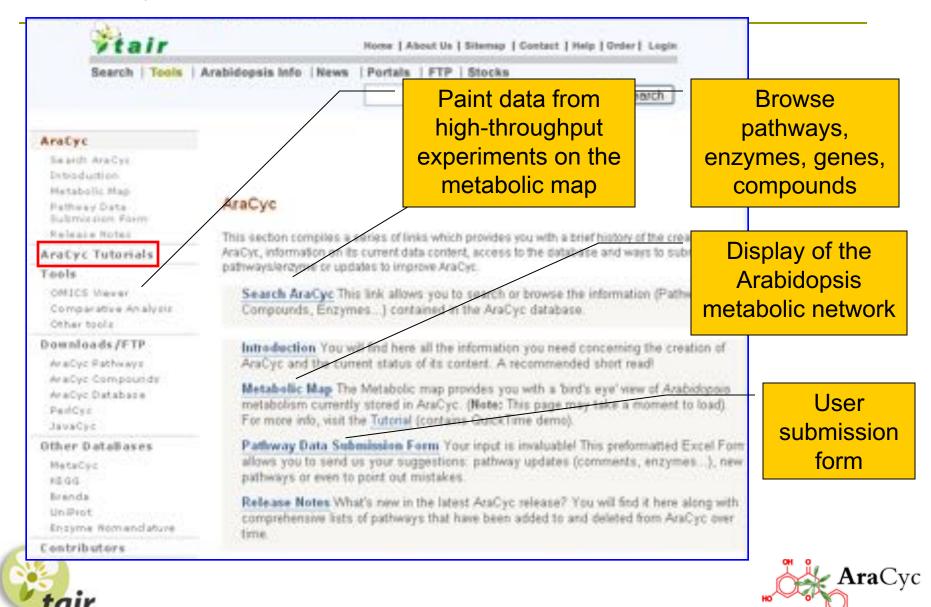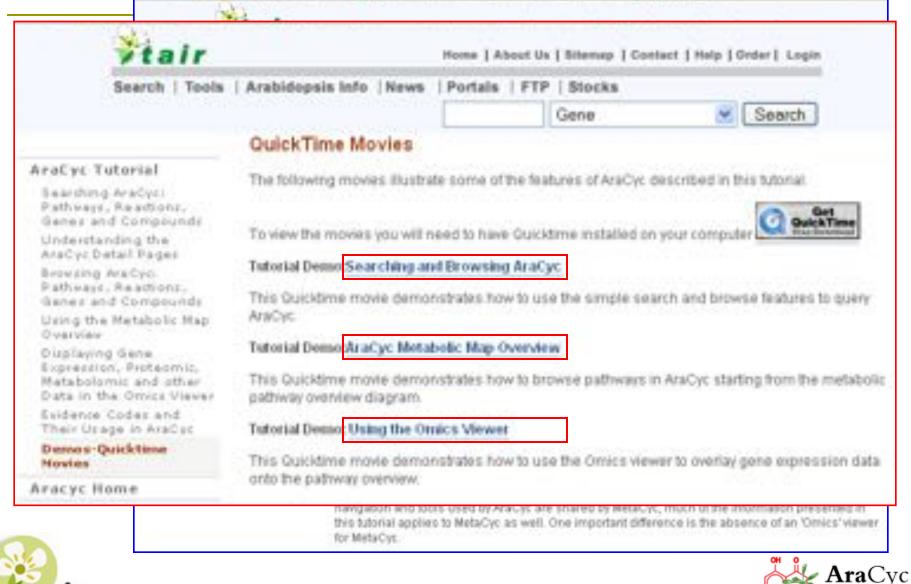
# AraCyc – The Home Page
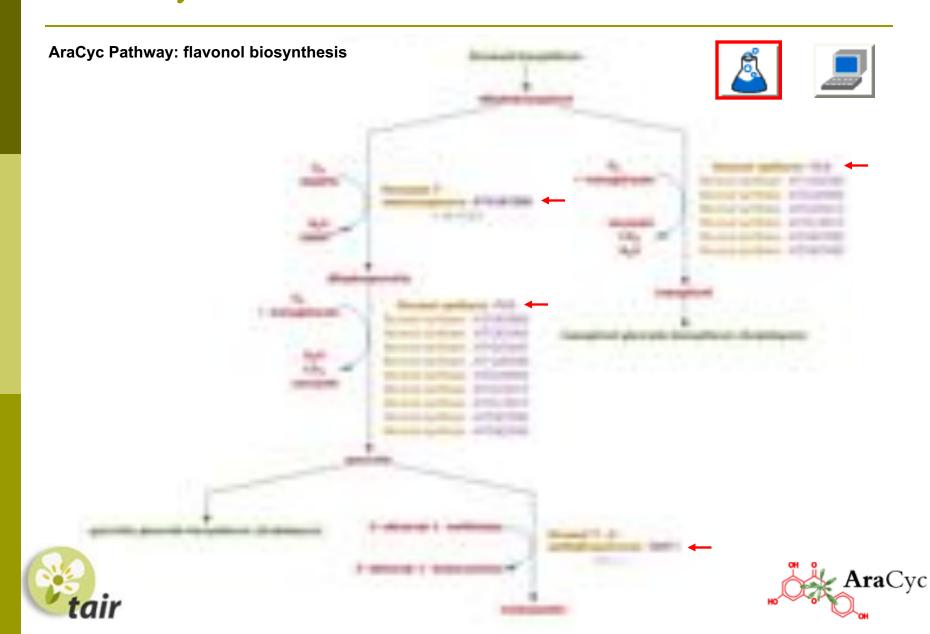
# AraCyc – All the Help you Can Get

# AraCyc's Content

**AraCyc Pathway: flavonol biosynthesis**

# Evidence Codes

Experimental E[vidence]ce:

Intuitive icons
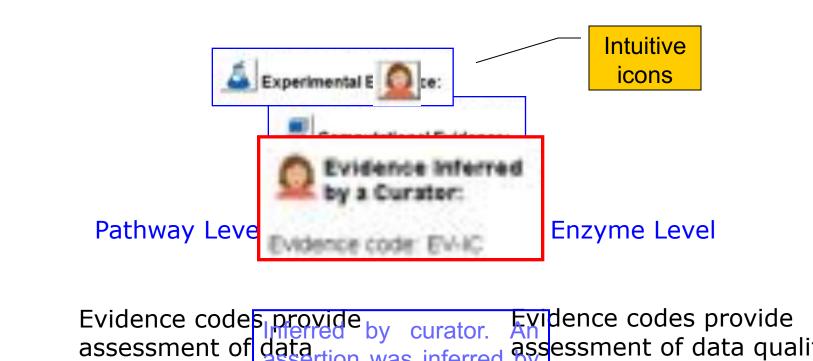
Evidence Inferred by a Curator:

Evidence code: EV-IC

Pathway Level

Enzyme Level

Evidence codes provide assessment of data quality, i.e. the affirmation for the existence of an pathway

Inferred by curator. An assertion was inferred by a curator from relevant information such as other assertions in a database

Evidence codes provide assessment of data quality, i.e. the affirmation for the catalytic activity of an enzyme

tair

AraCyc

# Evidence Codes *(cont'd)*

# AraCyc: Pathway Detail Page

# AraCyc: Metabolic Map



**AraCyc**

Search AraCyc
Introduction
Metabolic Map
Pathway Data Submission Form
Release Notes

**AraCyc Tutorials**

**Tools**

OMSCS Viewer
Comparative Analysis
Other tools

**Downloads/FTP**

AraCyc Pathways
AraCyc Compounds
AraCyc Database
PerlCyc
JavaCyc

**Other DataBases**

MetaCyc
KEGG
Brenda
UniProt
Enzyme Nomenclature

**Contributors**

## AraCyc

This section compiles a series of links which provides you with a brief history of the creation of AraCyc, information on its current data content, access to the database and ways to submit new pathways/enzyme or updates to improve AraCyc.

**Search AraCyc** This link allows you to search or browse the information (Pathways, Compounds, Enzymes...) contained in the AraCyc database.

**Introduction** You will find here all the information you need concerning the creation of AraCyc and the current status of its content. A recommended short read!

**Metabolic Map** The Metabolic map provides you with a 'bird's eye' view of Arabidopsis metabolism currently stored in AraCyc. (**Note**: This page may take a moment to load). For more info, visit the Tutorial (contains QuickTime demo)
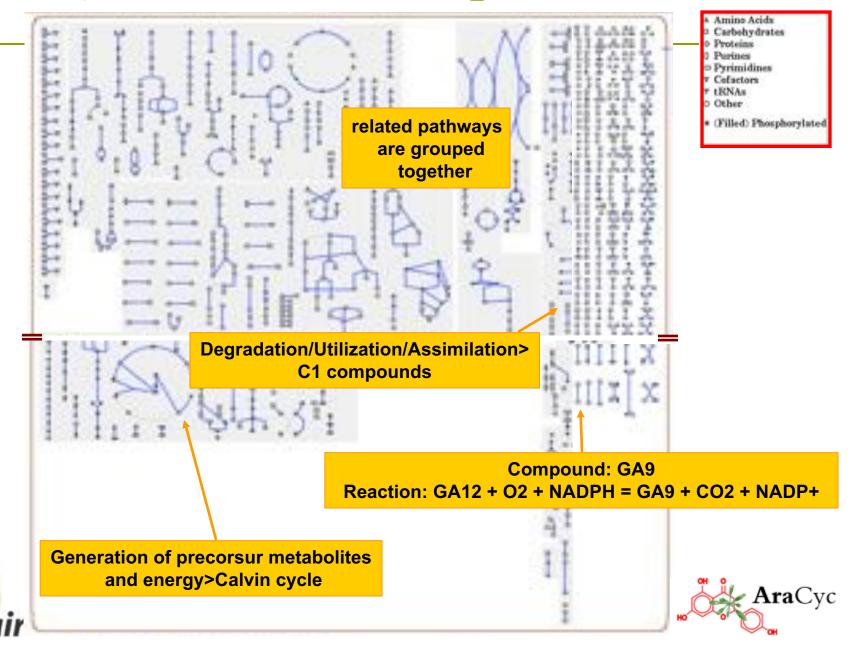
**Pathway Data Submission Form** Your input is invaluable! This preformatted Excel Form allows you to send us your suggestions: pathway updates (comments, enzymes...), new pathways or even to point out mistakes.

**Release Notes** What's new in the latest AraCyc release? You will find it here along with comprehensive lists of pathways that have been added to and deleted from AraCyc over time.

Troubleshooting

# AraCyc: Metabolic Map (*cont'd*)



Legend:
- Amino Acids
- Carbohydrates
- Proteins
- Purines
- Pyrimidines
- Cofactors
- tRNAs
- Other
- (Filled) Phosphorylated

related pathways are grouped together

Degradation/Utilization/Assimilation> C1 compounds

Compound: GA9
Reaction: GA12 + O2 + NADPH = GA9 + CO2 + NADP+

Generation of precorsur metabolites and energy>Calvin cycle

# AraCyc: OmicsViewer

# OMICS Viewer

- Part of the Pathway Tools Software Suite

- Displays bird-eye view of the Metabolic Overview diagram for a single organism
  - KEGG pathways are 'superpathways' without consideration of species specificity and pathway variants

- Allows to paint data values from the user's high-throughput onto the Metabolic Overview diagram

  - **Microarray Expression Data**
  - **Proteomics Data**
  - **Metabolomics Data**

AraCyc

# OmicsViewer Submission Page

## Pathway Tools Omics Viewer

**Step 1**

The Pathway Tools Omics Viewer (formerly the Pathway Tools Expression Viewer) paints data values from the user's high-throughput and other experiments onto the Metabolic Overview diagram for an organism.

The Omics Viewer can be used for:

- **Microarray Expression Data:** Reaction lines (and protein icons, where present) are color-coded according to the relative or absolute expression level of the gene that codes for the enzyme that catalyzes that reaction step. The Omics Viewer allows a scientist to interpret the results of gene-expression experiments in a pathway context.
- **Proteomics Data:** Reaction lines (and protein icons, where present) are color-

Select a dataset: Arabidopsis thaliana CO

File containing experimental data (NOT a URL): C:\Documents and Setti   Browse...

Do you want to display absolute or relative data values? Relative

If displa values,

Data va

The ite (zeroth datafile

Note: E

Load sample file and provide information about your data

**Step 2**

| Arabidopsis locus id | time_point_1 | time_point_2 | time_point_3 | time_point_4 |
|---|---|---|---|---|
| At1g77760 | 1.15 | 2.3 | 3.2 | 2.15 |
| At2g13360 | 0.7 | -0.53 | 0 | -0.73 |
| At3g10230 | -1.1 | -0.05 | 1.06 | 1.15 |
| At3g10230 | -0.65 | -0.58 | 1.13 | 1.23 |
| At3g01120 | -1.08 | -0.15 | -1.2 | -1.15 |
| At3g01500 | 0.07 | -0.72 | -0.68 | 1.4 |
| At3g02470 | 0.03 | -0.53 | 0.58 | 1.28 |
| At3g02470 | 0.55 | -0.12 | 0.62 | 0.65 |
| At3g02580 | 0.6 | -0.55 | 0.08 | 0.55 |
| At3g02580 | 1.15 | 0.7 | 0.03 | -0.6 |
| At3g02780 | -1.15 | 0.05 | 0.1 | -0.08 |
| At3g04120 | -0.15 | -1.55 | 0.12 | -0.3 |
| At3g04120 | -0.15 | -1.5 | 0.05 | -0.32 |

**Sample data file (text tab-delimited)**

0   1   2   3   4

AraCyc

tair

# OmicsViewer Submission Page (*cont'd*)



The Pathway Tools Omics Viewer (formerly the Pathway Tools Expression Viewer) paints data values from the user's high-throughput and other experiments onto the Metabolic Overview diagram for an organism.

The Omics Viewer can be used for:

- **Microarray Expression Data:** Reaction lines (and protein icons, where present) are color-coded according to the relative or absolute expression level of the gene that codes for the enzyme that catalyzes that reaction step. The Omics Viewer allows a scientist to interpret the results of gene-expression experiments in a pathway context.
- **Proteomics Data:** Reaction lines (and protein icons, where present) are color-coded according to the concentration of the enzyme that catalyzes that reaction step.
- **Metabolomics Data:** Compound icons

Select a dataset: Arabidopsis thaliana

File containing experimental data (NOT a URL): C:\Documents and Set

Do you want to display absolute or relative data values? Relative

**Step 3**

If displaying relative data values, use
- ⦿ a single data column
- ○ the ratio of two data columns

**Step 4**

Data values use a:
- ⦿ 0-centered scale (e.g. log scale)
- ○ 1-centered scale (negative values will be discarded)

**Step 5**

The items in the first (zeroth) column of your datafile are: Genes

Genes
Proteins
Compounds
Reactions
Any of the above

**Step 6**

Note: By selecting Any of the above combine, for example, gene expression and metabolomics display. There are some dangers inherent in this ambiguous if it is not known names may be proteins or metabolites. In addition, data values from different kinds of experiments may not be directly

Choose to display a single or multiple

Select the type of data you want to display (refers to your loading file)

**Ara**Cyc

# OmicsViewer Submission Page (*cont'd*)

**0**   **1**   **2**   **3**   **4**

| Arabidopsis locus id | time_point_1 | time_point_2 | time_point_3 | time_point_4 |
|---|---|---|---|---|
| At1g77760 | 1.15 | 2.3 | 3.2 | 2.15 |
| At2g13360 | 0.7 | -0.53 | 0 | -0.73 |
| At3g10230 | -1.1 | -0.05 | 1.05 | 1.15 |

For single/multiple or the ratio of time points add the corresponding column number(s)

assigned to genes, proteins, reactions or metabolites can be viewed in a pathway context using the Omics Viewer.

More information about the Omics Viewer, including sample datafiles and displays.

The Omics Viewer takes as input a tab-delimited data file that is stored on your local computer. The file contains relative or absolute data for a set of genes, proteins, reactions, metabolites, or some combination. Each row of the file contains data for a single gene, protein, reaction, or metabolite, and begins with the object name, ID or EC number. You may choose to display either data from a single column, the ratio of two columns, or a time series animation of multiple columns. Data

**Time Step or Ani...**

...ment time step, ente... ...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):

```
2
3
4
```

If using two columns, denominator data column:

```
1
1
2
```

Step 7

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.

tair

AraCyc

# OmicsViewer Submission Page (*cont'd*)

Choose your cutoff to visualize your expression values

Choose a color scheme:
- ⦿ Full color spectrum, computed from data provided (default)
- ○ Full color spectrum with a maximum cutoff: ▭
- ○ Three color display with specified threshhold: ▭

**Step 8**

## Display Type

By default, data values are painted on the cellular overview chart. However an alternative display is to generate a table containing all individual pathways which have one or more data values that exceed some threshhold (or are less than the inverse of that threshhold). To select this alternative display, choose the corresponding option below and specify the threshhold.

- ⦿ Paint data on overview chart (default)
- ○ Generate a table of individual pathways exceeding threshhold: ▭
- ○ Combine both displays (not yet implemented for animations)

**Step 9**

Submit  Note that this request will take several minutes to complete (p... longer for large datasets).

*tair*

**Ara**Cyc

# The Omics Viewer Result Page



compounds (icons) are color-coded according to the concentration of compounds

reactions (lines) are color-coded according to the gene expression level

Amino Acids
Carbohydrates
Proteins
Purines
Pyrimidines
Cofactors
tRNAs
Other

(Filled) Phosphorylated

Expression Ratios
50.
19.
7.
2.5
1.
0.4
0.15
6.00e-2
2.00e-2

No data available
Unknown enzyme

Histogram

26
0 3
left side =
  x in overview
right side =
  x not in overview
Multiple highlights

# The Omics Viewer Result Page *(cont'd)*



- Carbohydrates
- Proteins
- Purines
- Pyrimidines
- Cofactors
- tRNAs
- Other

- (Filled) Phosphorylated

**Expression Ratios**

+4.9
+3.7
+2.5
+1.3
0.
−1.3

The statistics for the expression map (single time points only) is provided at the bottom of the page

Total number of data rows (not including comment lines): 493
Number of rows for which the gene could not be found: 1
Number of rows for which the gene name was ambiguous: 4
Number of rows for which the gene is valid, but for which a data value was missing or malformed:

The following table shows statistics for the selected column/column ratio for all genes and for those

| Data Statistics | All Genes | Overview Genes |
|---|---|---|
| Number of values: | 487 | 463 |
| Minimum value: | -4.55 | -4.55 |
| Maximum value: | 7.85 | 7.85 |
| Median: | -0.51 | -0.53 |
| Mean: | -0.11884997 | -0.1369763 |
| Standard deviation: | 1.4973682 | 1.5045344 |

Objects that could not be found: At2g18440
Ambiguous object names: is ambiguous for G1F4, G1F0 is ambiguous for G1F4, G1F0 is ambiguous for G1F4, G1F0 is ambiguous for G1F4, G1F0
Objects with missing or malformed data: At1g06460

tair

vailable
enzyme

t side =
enes in overview
ght side =
enes not in overview
ultiple highlights

**Ara**Cyc

# Saving Results

# Acknowledgements

□ **TAIR**
- Sue Rhee *(PI)*
- Peifen Zhang *(curator)*
- Christophe Tissier (*curator*)
- Hartmut Foerster *(curator)*
- Tom Walk (post doctoral researcher)

□ **SRI**
- Peter Karp *(PI)*
- Ron Caspi *(curator)*
- Carol Fulcher *(curator)*
- Suzanne Paley *(software developer)*
- Pallavi Kaipa *(programmer)*
- Markus Krummenacker *(programmer)*

□ **Previous contributers**
- Lukas Müller
- Aleksey Kleytman *(curator assistant – TAIR)*
- Thomas Yan *(programmer – TAIR)*
- Joe Filla *(sysadmin - TAIR)*
- Mary Montoya *(software developer - NCBR)*
- John Pick *(software developer - SRI)*
- Mario Latendresse *(programmer - SRI)*

*NIH,     NSF,     Pioneer Hi-Bred*

**tair**   **MetaCyc**   **AraCyc**