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

Currently available PGDBs

PGDB	Species	Host	Status
AraCyc	Arabidopsis	TAIR/PMN	Substantial curation
RiceCyc	Rice	Gramene	Some curation
	Sorghum	Gramene	No curation
MedicCyc	Medicago	Noble Foundation	some curation
LycoCyc	Tomato	SGN	some curation
	Potato	SGN	No curation
	Pepper	SGN	No curation
	Tobacco	SGN	No curation
	Petunia	SGN	No curation
	Coffee	SGN	No curation

Plant Metabolic Pathway Databases

PlantCyc Search More Info

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

AraCyc Search More Info Overlay my data

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 annuum*)
- CoffeaCyc (Coffee) (*Coffea canephora*)
- LycoCyc (Tomato) (*Solanum lycopersicum*)
- MedicCyc (Medicago) (*Medicago truncatula*)
- NicotianaCyc (Tobacco) (*Nicotiana tabacum*)
- PetuniaCyc (Petunia) (*Petunia hybrida*)
- 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#) caffeine Submit
  - 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 c 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. 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**

Pathways Pathway pages contain: Depiction of metabolic pathway,

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

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

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

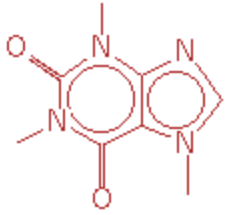
Compounds Compound pages contain: compound structural inform

- caffeine

Synonyms: 1,3,7-trimethylxanthine

Empirical Formula: $C_8H_{10}N_4O_2$

Molecular Weight: 194.19 daltons

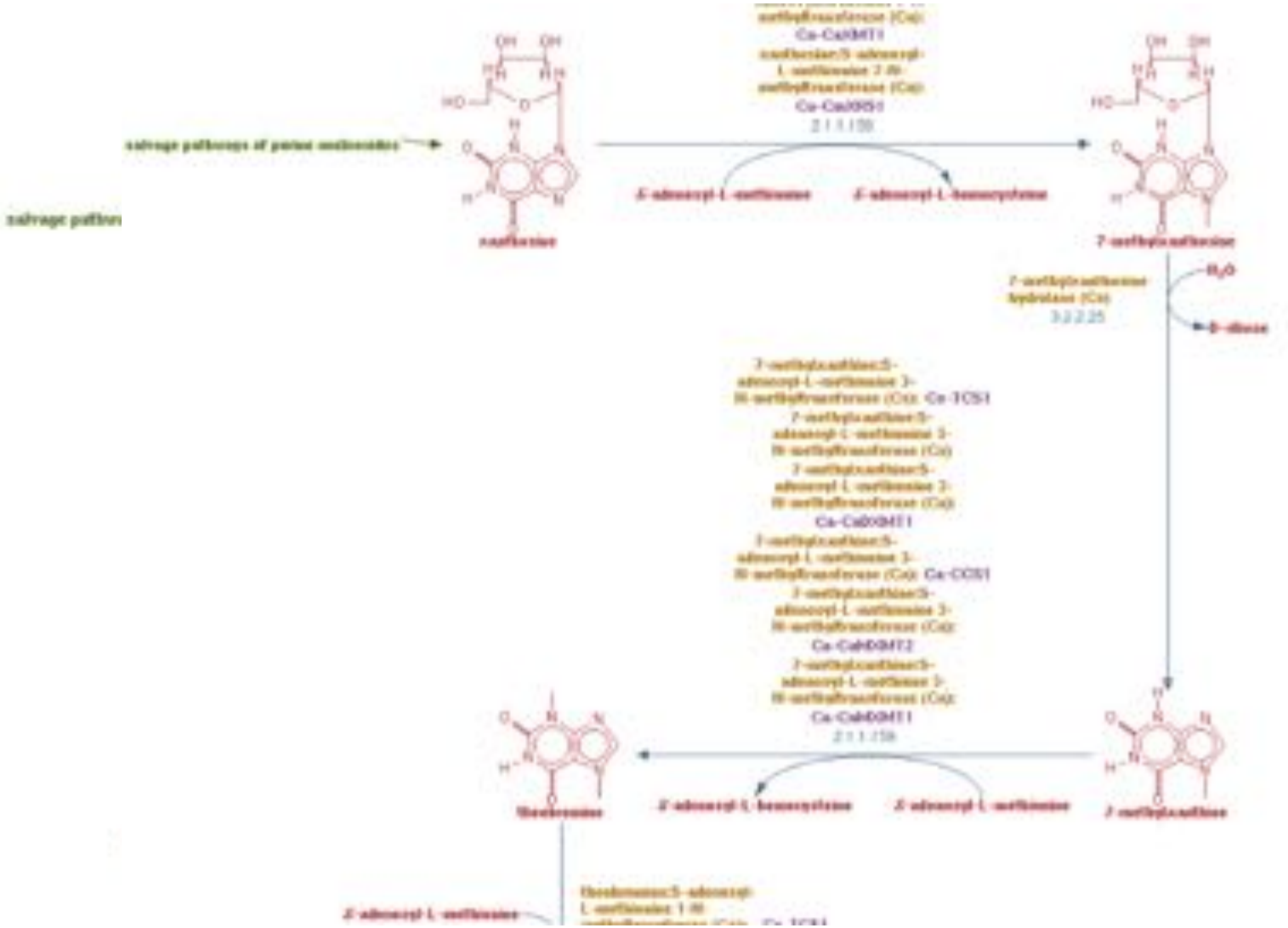


Smiles: c12(n(C)c(=O)n(C)c(=O)c(n(C)cn1)2)

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



Superclasses: Biosynthesis -> Secondary Metabolites Biosynthesis -> Plant Secondary Metabolites Biosynthesis -> Nitrogen-Containing Secondary Compounds Biosynth
Alkaloids Biosynthesis -> Purine alkaloids -> Caffeine Biosynthesis

Species Data Available for: Camellia sinensis , Camellia sinensis assamica , Camellia taliens

Summary:

Caffeine, 1,3,7-trimethylxanthine, is one of the best known purine alkaloids and also the mo recurs
caffeine (3,7-dimethylxanthine) [Ashihara04]. Caffeine presents in high concentrations in r ht in s
young expanding leaves of Coffea arabica, and up to 3% dry weight in young leaves of tea (e alka
plants is undetermined. Their hypothetic roles include being a defense chemical where hig ower l
protect these soft tissues from predators, or being an autotoxic chemical where caffeine rel s. Hov
is little evidence supports the hypothesis.

The depicted pathway herein illustrates the major route of caffeine biosynthesis, from xanth name
xanthosine N-methyltransferase, 7-methylxanthine N-methyltransferase, and theobromine n hav
specific xanthosine N-methyltransferase activity. It specifically converts xanthosine to 7-met rate.
Theobromine synthase has also been characterized from coffee which has the specific 7-m e to th
(the second methylation step). Caffeine synthase has been characterized from both coffee meth
N-methyltransferase and the theobromine N-methyltransferase activities. It converts 7-meth d the
methylation steps. Immature fruits are the major sites of caffeine biosynthesis in coffee pla ltrans
activities are also found [Uefuji03].

Free purine nucleotides are the major resources of xanthosine. Xanthosine is derived from of pur
nucleosides). Newly synthesized IMP from de novo purine biosynthesis, and adenosine re also
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 s II (vi
paraxanthine)).

Citations: [Misako04 , Ogawa01 , Mizuno03b , Mizuno03a , Kato96]

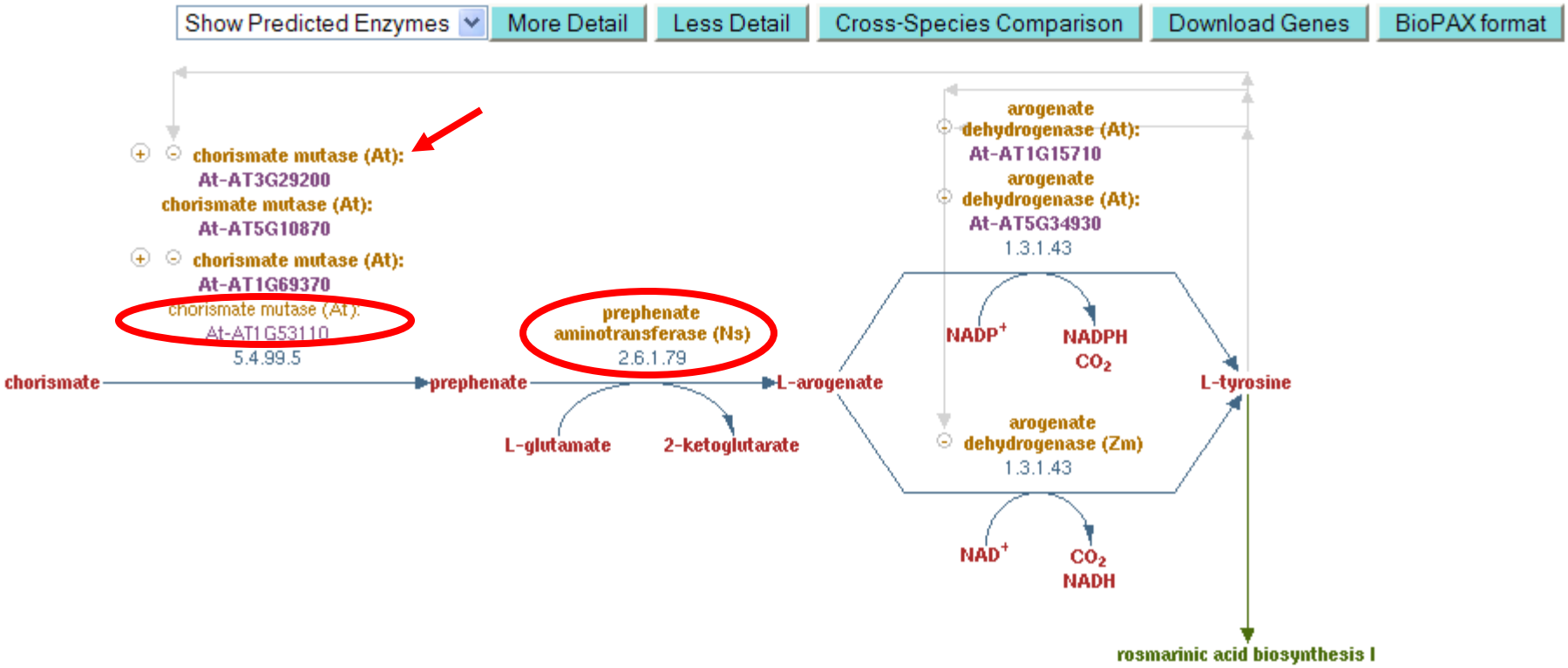
Variants: caffeine biosynthesis II (via paraxanthine)

Parent Classes:
- Nitrogen-Containing Secondary Compounds Biosynthesis

Child Classes:
- Betalaine alkaloids (8)
- Indole alkaloids (5)
- Isoquinoline and Benzylisoquinoline alkaloids (8)
- Peptide alkaloids (0)
- Polyketide alkaloids (0)
- Purine alkaloids (4)
- Pyrrolidine, Piperidine and Pyridine alkaloids (3)
- Pyrrolizidine alkaloids (0)
- Quinoline alkaloids (0)
- Quinolizidine alkaloids (1)
- Terpenoid Alkaloids Biosynthesis (3)
- Tropane alkaloids (3)

Instances:
- capsaicin biosynthesis
- ephedrine biosynthesis
- γ-coniciene and coniine biosynthesis
- N-methyl-Δ^1-pyrrolinium cation biosynthesis
- steroidal glycoalkaloid biosynthesis
- tryptophan degradation (via tryptamine)

PlantCyc Pathway: tyrosine biosynthesis I



PlantCyc Enzyme: chorismate mutase / L-ascorbate peroxidase

Enzymatic reaction of: chorismate mutase

chorismate <=> prephenate

The reaction direction shown, that is, A + B <==> C + D versus C + D <==> A + B, is

Reversibility of this reaction is unspecified.

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

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

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

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

K_M for chorismate: 2900 µM [Mobley99]

Enzymatic reaction of: L-ascorbate peroxidase

2 L-ascorbate + H_2O_2 <=> 2 monodehydroascorbate + 2 H_2O

The reaction direction shown, that is, A + B <==> C + D versus C + D <==> A + B, is

Reversibility of this reaction is unspecified.

In Pathways: ascorbate glutathione cycle

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

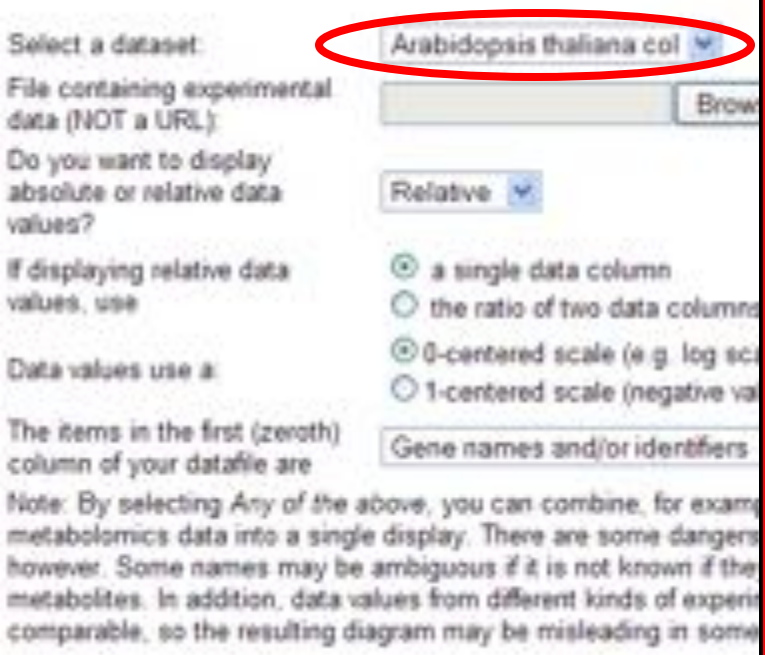
Omics Viewer

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



At1g77760	1.15	2.3	3.2	2.15	1.53
At2g13360	0.7	-0.53	0	-0.73	0.03
At3g10230	-1.1	-0.05	1.05	1.15	1.25
At3g10230	-0.65	-0.58	1.13	1.23	0.67
At3g01120	-1.08	-0.15	-1.2	-1.15	-1.15
At3g01500	0.07	-0.72	-0.68	-1.4	-1.93
At3g02470	0.03	-0.53	0.58	1.28	0.55
At3g02470	0.55	-0.12	0.62	0.65	-0.05
At3g02580	0.6	-0.55	0.08	0.55	-2.2
At3g02580	1.15	0.7	0.03	-0.6	-2.4
At3g02780	-1.15	0.05	0.1	-0.08	-0.57
At3g04120	-0.15	-1.55	0.12	-0.3	0.23
At3g04120	-0.15	-1.5	0.05	-0.32	0.25
At3g04120	-0.07	-0.85	0.1	-0.75	0.2
At3g04870	1.05	-1.08	-0.05	-1.1	0.05
At3g04940	-0.85	-0.1	-0.85	-1.3	-1.83
At3g07420	-0.68	-0.12	-0.7	-0.1	0
At3g10850	-0.6	-0.78	-0.65	-0.72	-2.08
At3g13790	-0.2	1.8	1.65	1.75	1.77

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):

1
3
5

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:

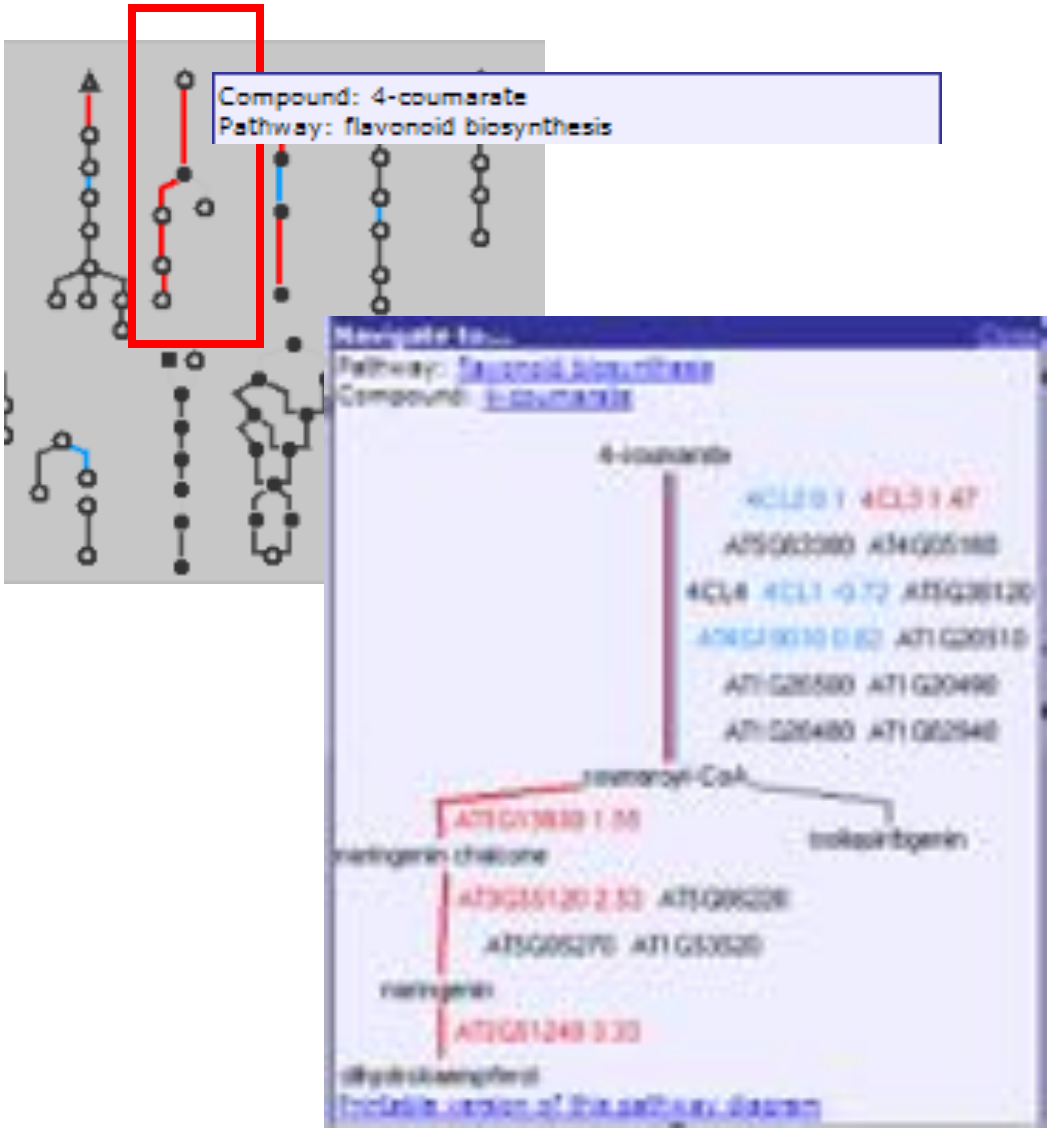
Submit Note that this request will take several minutes to complete (possibly longer for large datasets). For faster operation, install Pathway Tools on your own computer! Click here for details.

Red: enhanced expression over my threshold

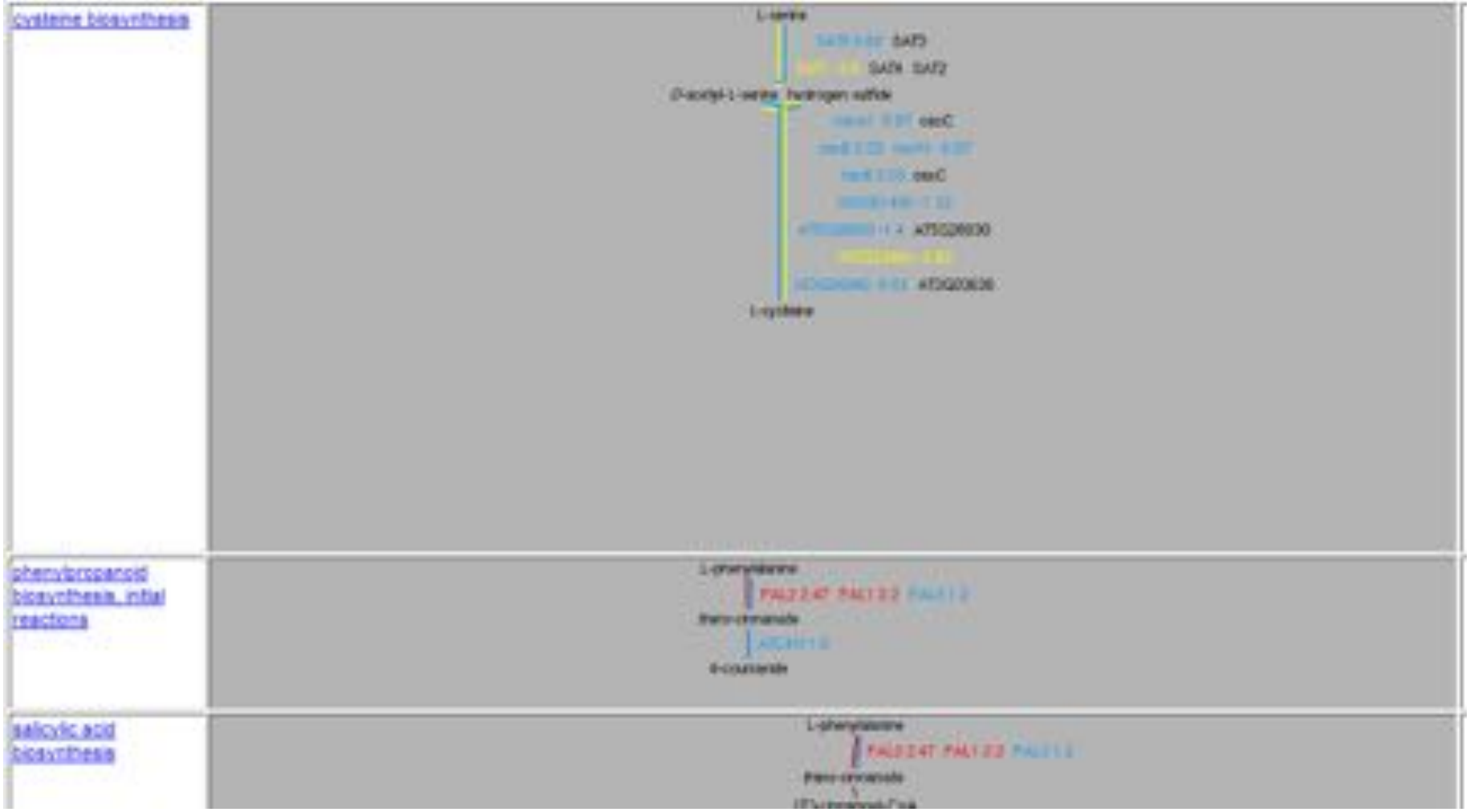
Yellow: repressed expression over my threshold

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

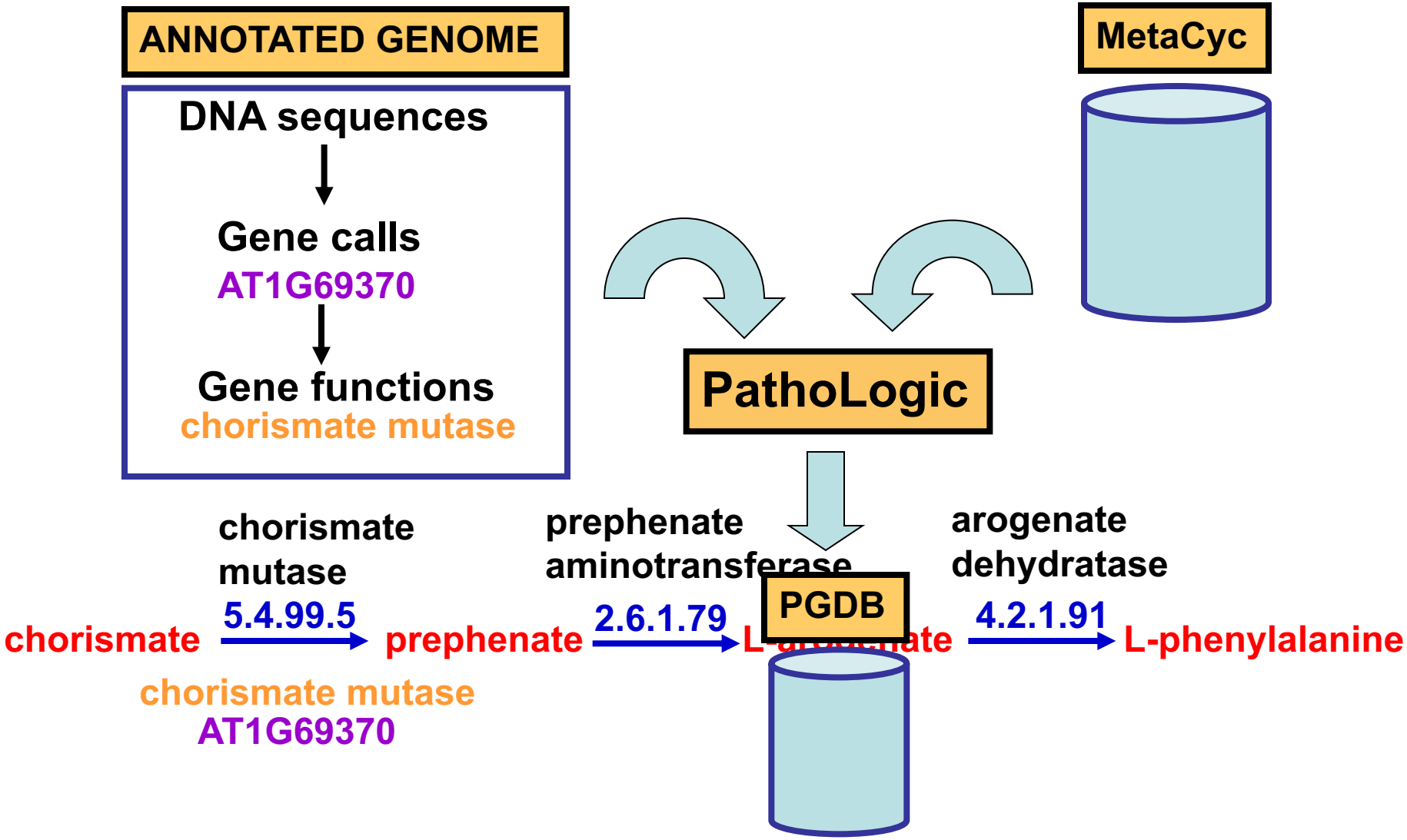
Create PGDBs, how

- 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)


ANNOTATED GENOME
DNA sequences
Gene calls
AT1G69370
Gene functions
chorismate mutase
MetaCyc
PathoLogic
PGDB
chorismate mutase
5.4.99.5
prephenate aminotransferase
2.6.1.79
arogenate dehydratase
4.2.1.91
chorismate
prephenate
L-phenylalanine
chorismate mutase
AT1G69370

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

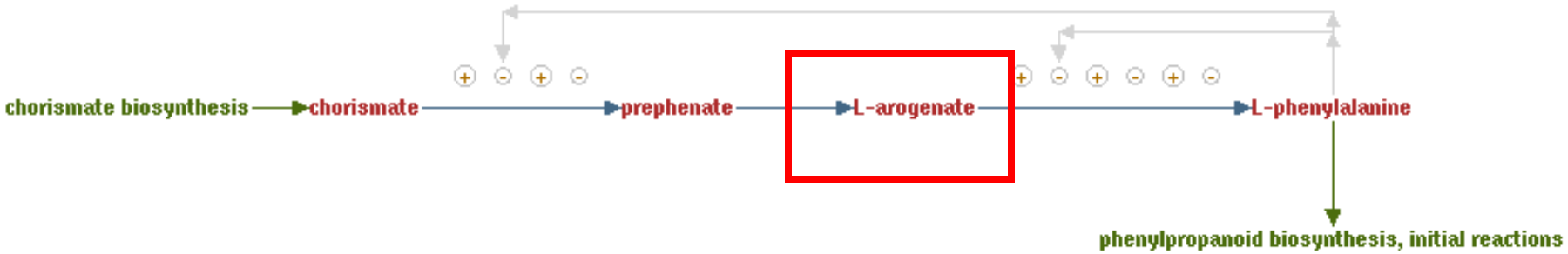
- Pathways operating via a different route
  - Phenylalanine biosynthesis in bacteria v.s. in plants

MetaCyc Pathway: phenylalanine biosynthesis I

More Detail | Less Detail | Cross-Species Comparison | BioPAX format

PlantCyc Pathway: phenylalanine biosynthesis

More Detail | Less Detail | Cross-Species Comparison | Download Genes | BioPAX format

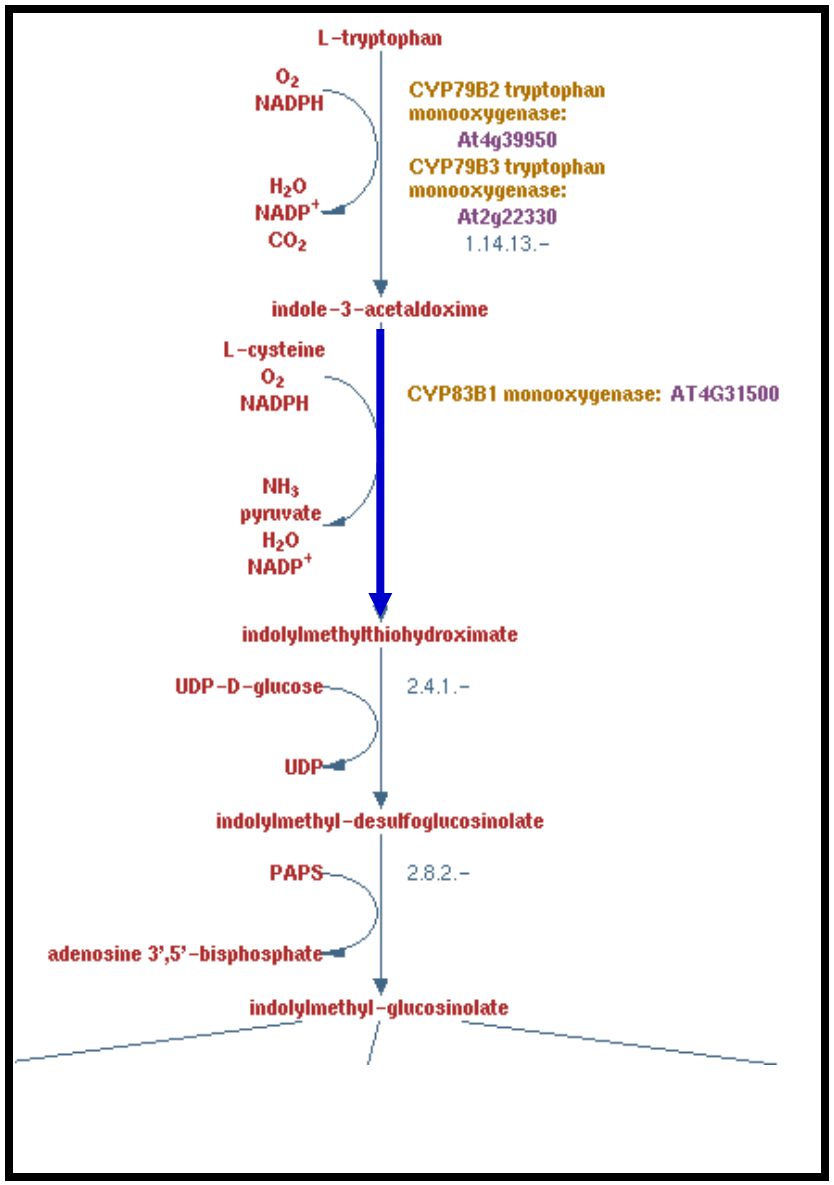


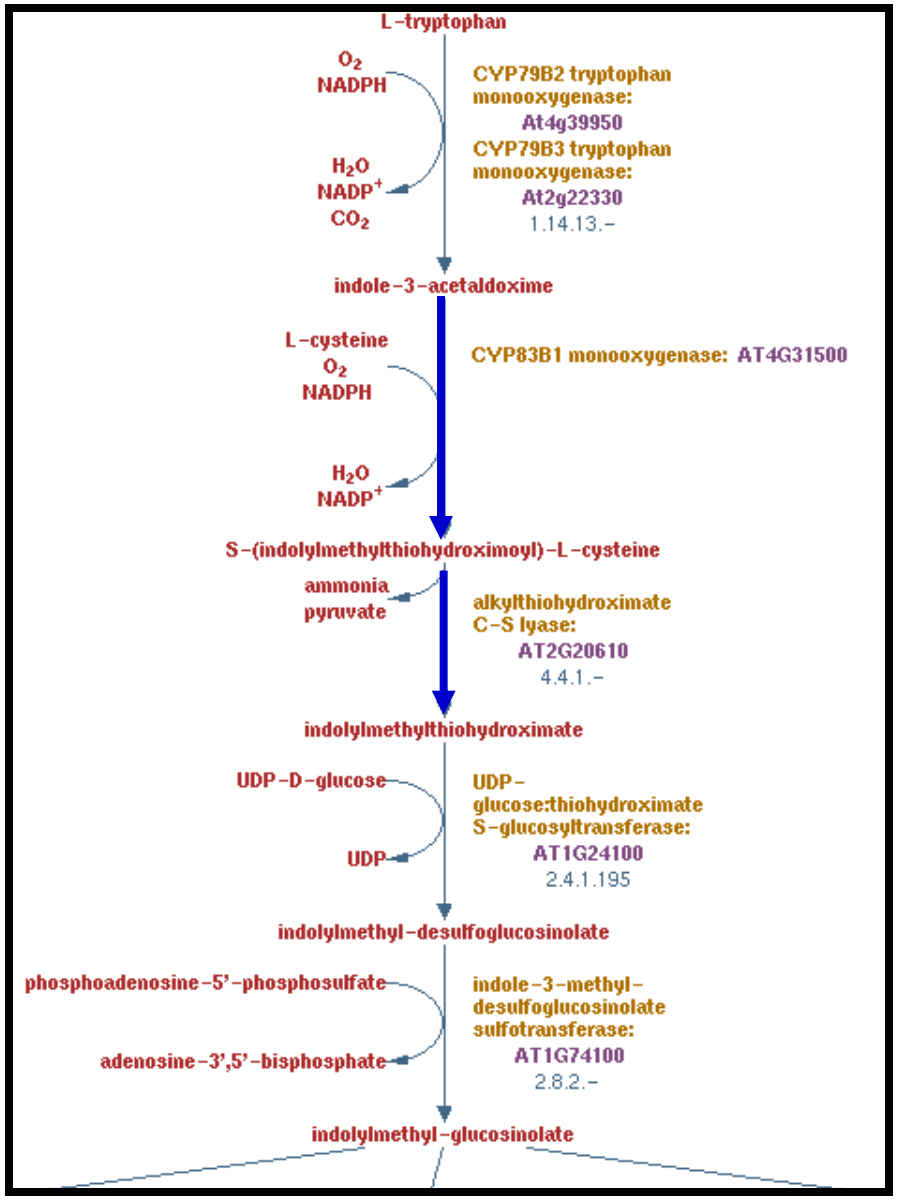
Validation: add evidence and literature support

***AraCyc* Pathway: phenylalanine biosynthesis**

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

Curation: correct pathway diagrams





Curation: correct gene/enzyme assignments to reaction/pathway

UGT89C1

Enzymatic reaction of: 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 <==> C + D versus C + D <==> A + B, is in accordance with t within a pathway.

Reversibility of this reaction is unspecified.

In Pathways: quercetin glucoside biosynthesis

Citations: [YonekuraSa07]

Enzymatic reaction of: 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!