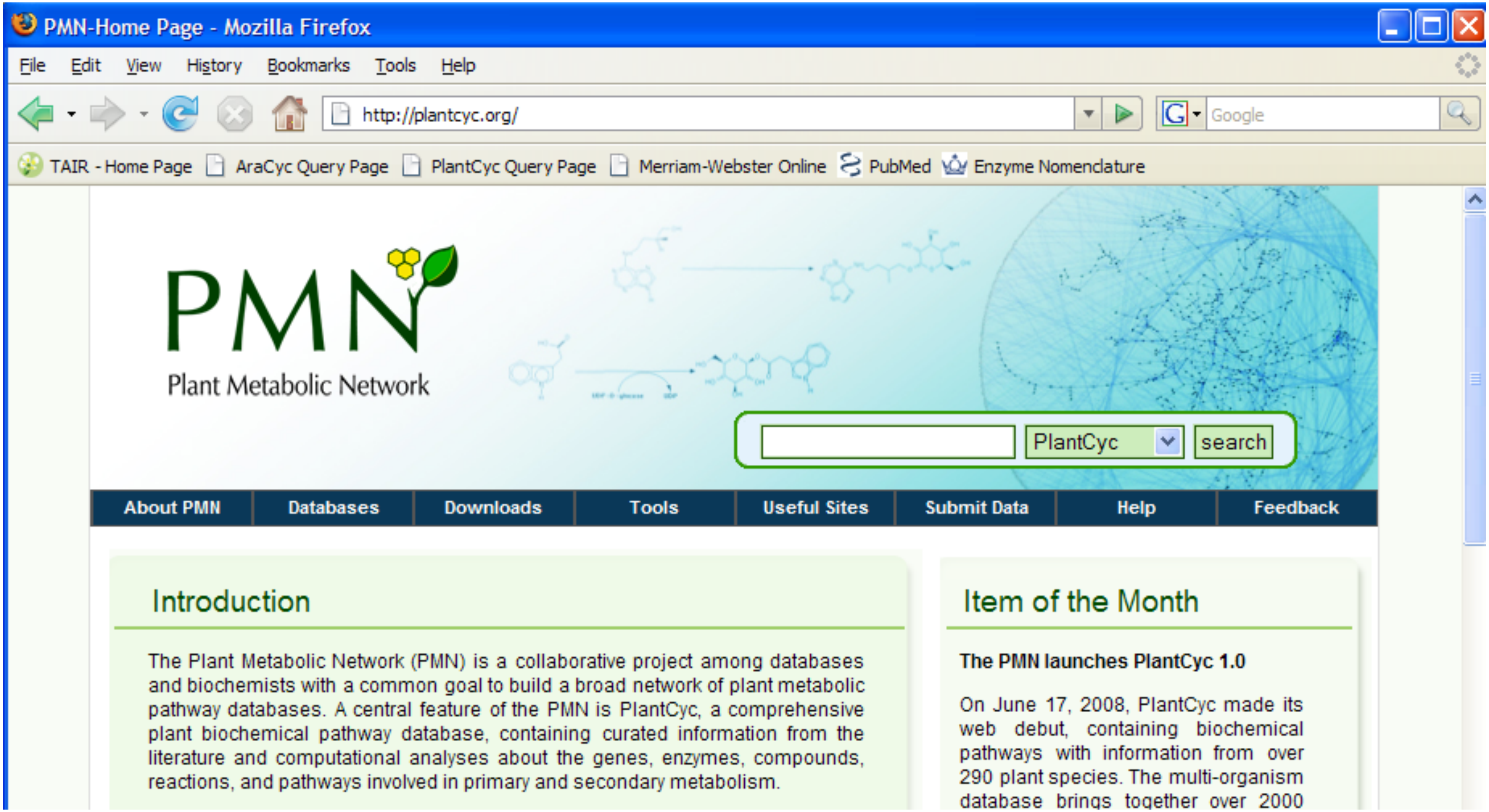
How pathway databases were created and curated

Peifen Zhang

Plant Metabolic Network (PMN)

About PMN, http://plantcyc.org



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

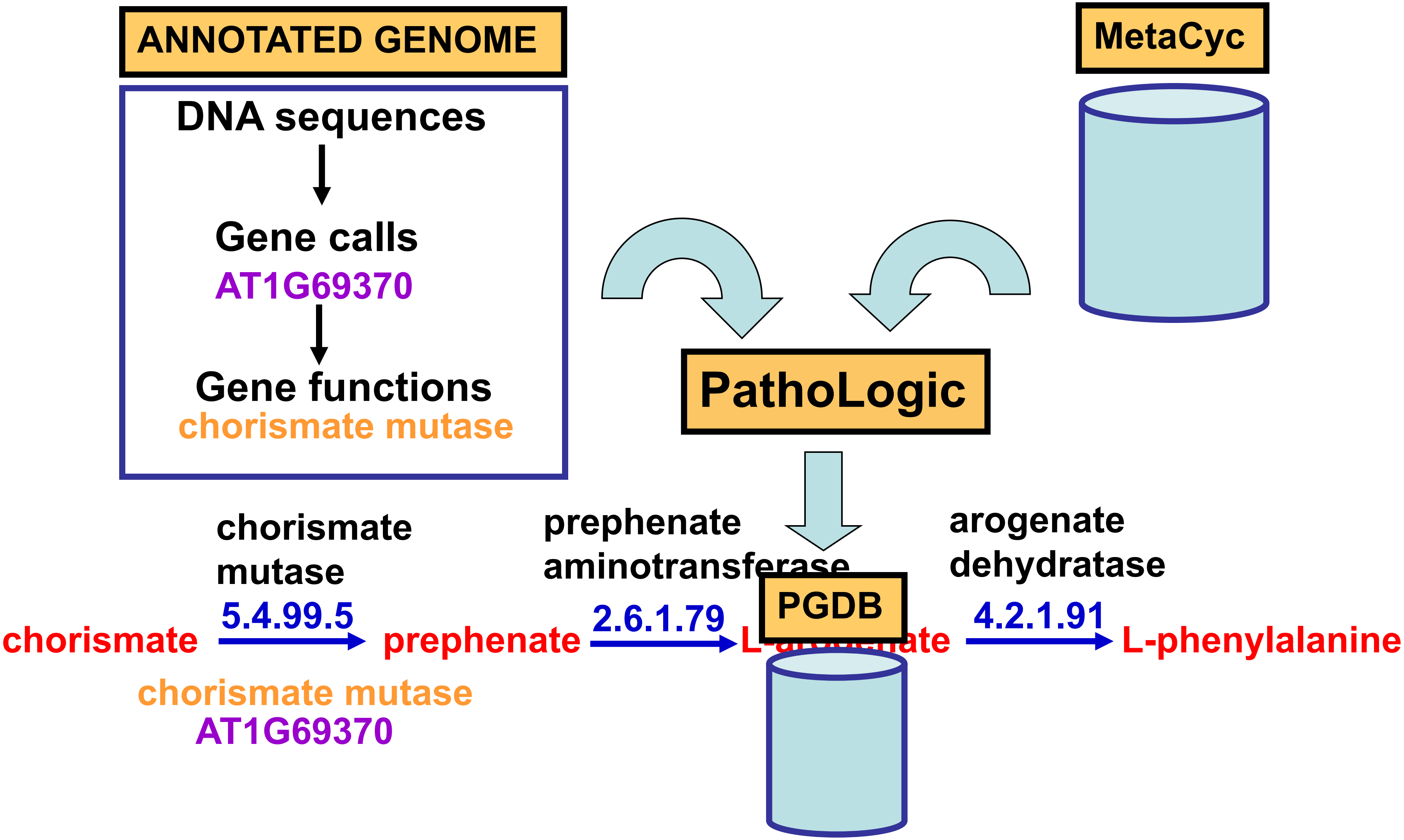
Creation of PGDBs, how

- 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)



ANNOTATED GENOME

DNA sequences

Gene calls

AT1G69370

Gene functions

chorismate mutase

MetaCyc

PathoLogic

PGDB

chorismate mutase

prephenate aminotransferase

arogenate dehydratase

chorismate $\xrightarrow{5.4.99.5}$ prephenate $\xrightarrow{2.6.1.79}$ L-arogenate $\xrightarrow{4.2.1.91}$ L-phenylalanine

chorismate mutase

AT1G69370

A snap shot 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

Species	Database	Status
Arabidopsis	TAIR	Substantial curation
Rice	Gramene	Some curation
Sorghum	Gramene	No curation
Medicago	Noble Foundation	some curation
Tomato	SGN	some curation
Potato	SGN	No curation
Pepper	SGN	No curation
Tobacco	SGN	No curation
Petunia	SGN	No curation
Coffee	SGN	No curation

Prediction of new PGDBs by PMN

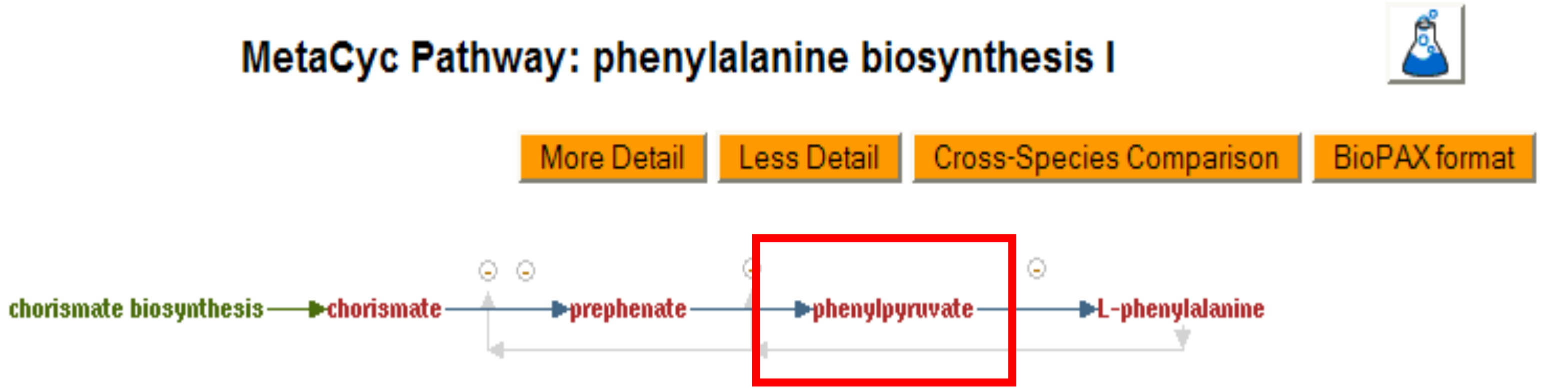
- Prioritization
  - Available sequences, economic impact
- High priority
  - Maize, Poplar, Soybean, Wheat
- Second priority
  - Cotton, Grape, Sugarcane, Sunflower, Switchgrass…

A quality database REQUIRES manual validation and curation

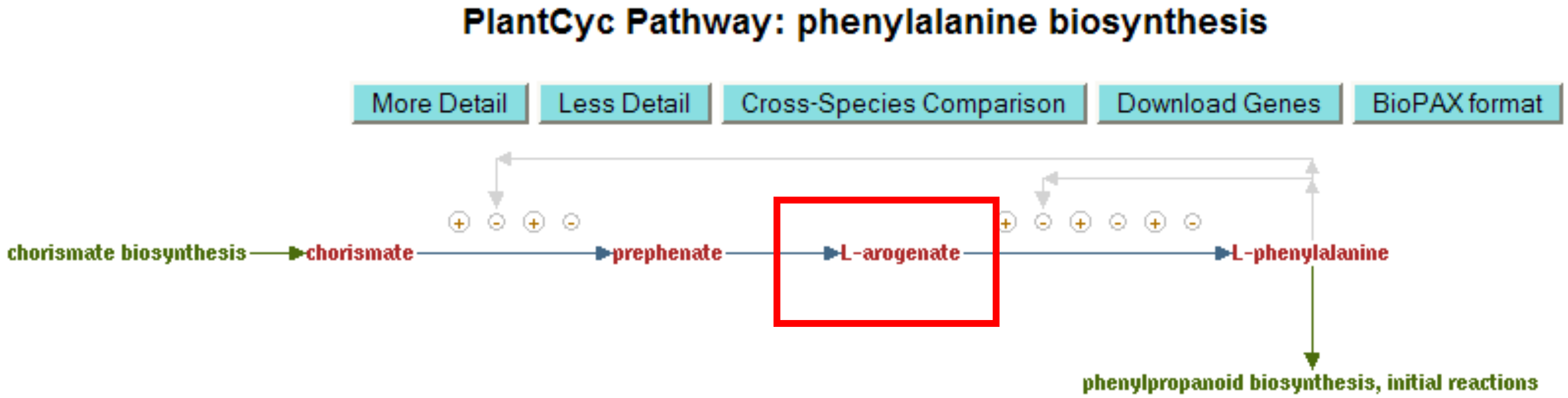
Validation: pruning false-positive predictions

- 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



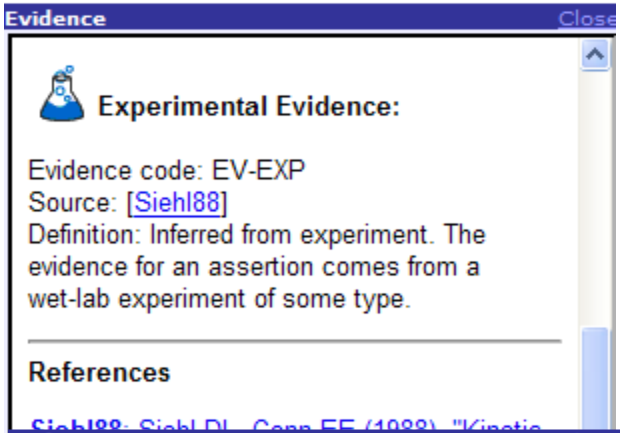
MetaCyc Pathway: phenylalanine biosynthesis I



PlantCyc Pathway: phenylalanine biosynthesis

Validation: adding evidence and literature supports

***AraCyc* Pathway: phenylalanine biosynthesis**



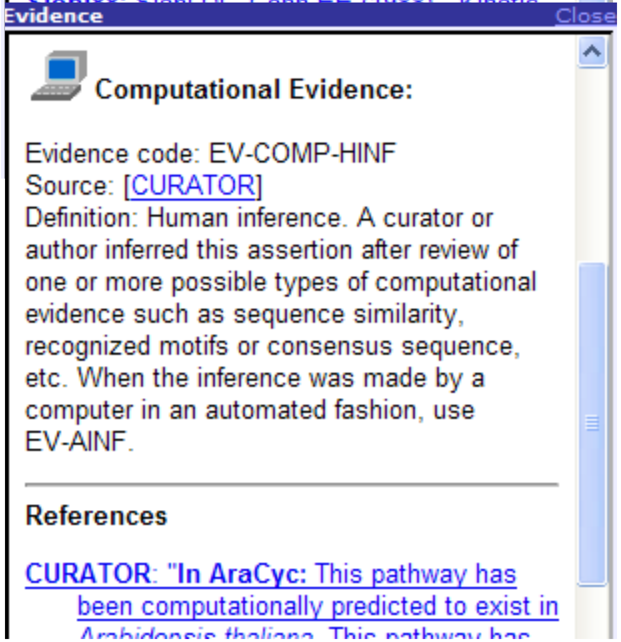
Experimental Evidence:

Evidence code: EV-EXP
Source: [Siehl88]
Definition: Inferred from experiment. The evidence for an assertion comes from a wet-lab experiment of some type.

References

Siehl88: Siehl DL, Conn EE (1988) "Kinetic

***AraCyc* Pathway: ribose degradation**



Computational Evidence:

Evidence code: EV-COMP-HINF
Source: [CURATOR]
Definition: Human inference. A curator or author inferred this assertion after review of one or more possible types of computational evidence such as sequence similarity, recognized motifs or consensus sequence, etc. When the inference was made by a computer in an automated fashion, use EV-AINF.

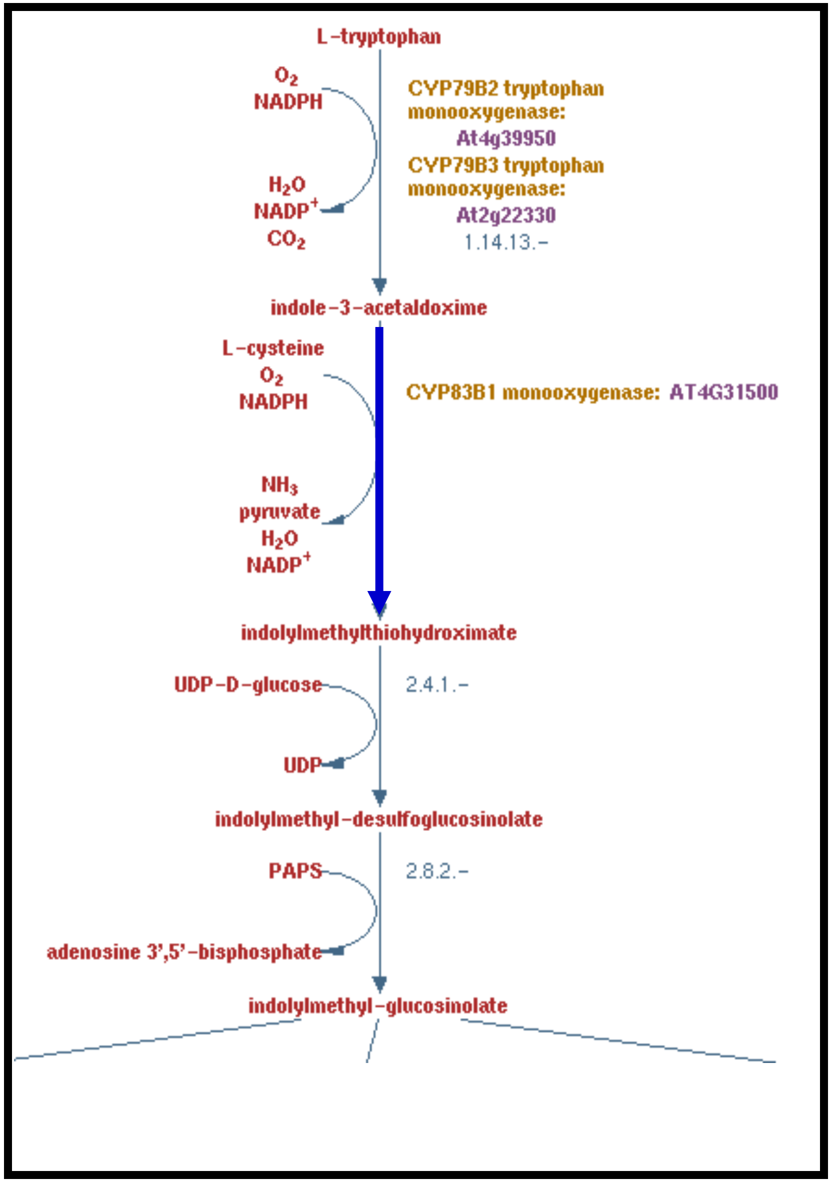
References

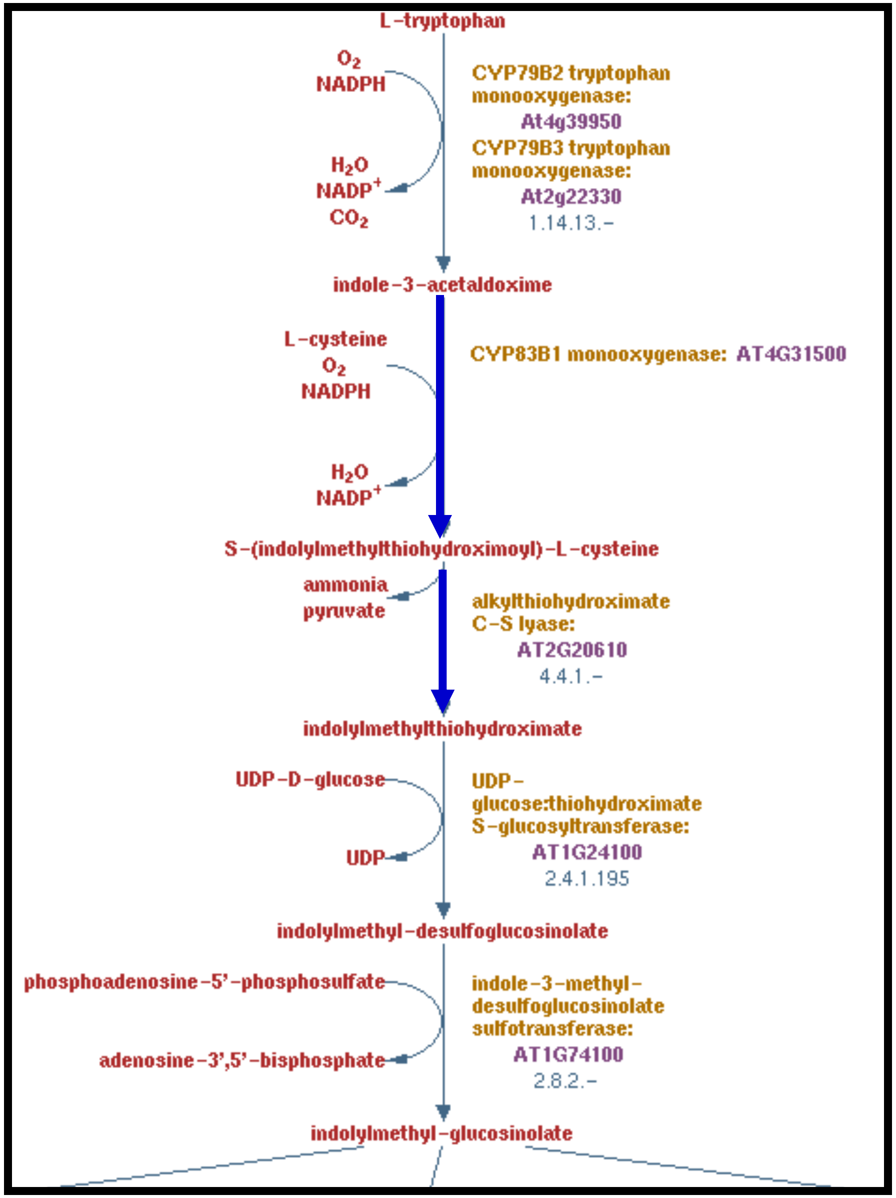
CURATOR: "**In AraCyc:** This pathway has been computationally predicted to exist in *Arabidopsis thaliana*. This pathway has

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


L-tryptophan
O_2
NADPH
H_2O
$NADP^+$
CO_2
CYP79B2 tryptophan monooxygenase:
At4g39950
CYP79B3 tryptophan monooxygenase:
At2g22330
1.14.13.-
indole-3-acetaldoxime
L-cysteine
O_2
NADPH
NH_3
pyruvate
H_2O
$NADP^+$
CYP83B1 monooxygenase: AT4G31500
indolylmethylthiohydroximate
UDP-D-glucose
UDP
2.4.1.-
indolylmethyl-desulfoglucosinolate
PAPS
adenosine 3',5'-bisphosphate
2.8.2.-
indolylmethyl-glucosinolate


L-tryptophan
O_2
NADPH
H_2O
$NADP^+$
CO_2
CYP79B2 tryptophan monooxygenase:
At4g39950
CYP79B3 tryptophan monooxygenase:
At2g22330
1.14.13.-
indole-3-acetaldoxime
L-cysteine
O_2
NADPH
H_2O
$NADP^+$
CYP83B1 monooxygenase: AT4G31500
S-(indolylmethylthiohydroximoyl)-L-cysteine
ammonia
pyruvate
alkylthiohydroximate C-S lyase:
AT2G20610
4.4.1.-
indolylmethylthiohydroximate
UDP-D-glucose
UDP
UDP-glucose:thiohydroximate S-glucosyltransferase:
AT1G24100
2.4.1.195
indolylmethyl-desulfoglucosinolate
phosphoadenosine-5'-phosphosulfate
adenosine-3',5'-bisphosphate
indole-3-methyl-desulfoglucosinolate sulfotransferase:
AT1G74100
2.8.2.-
indolylmethyl-glucosinolate

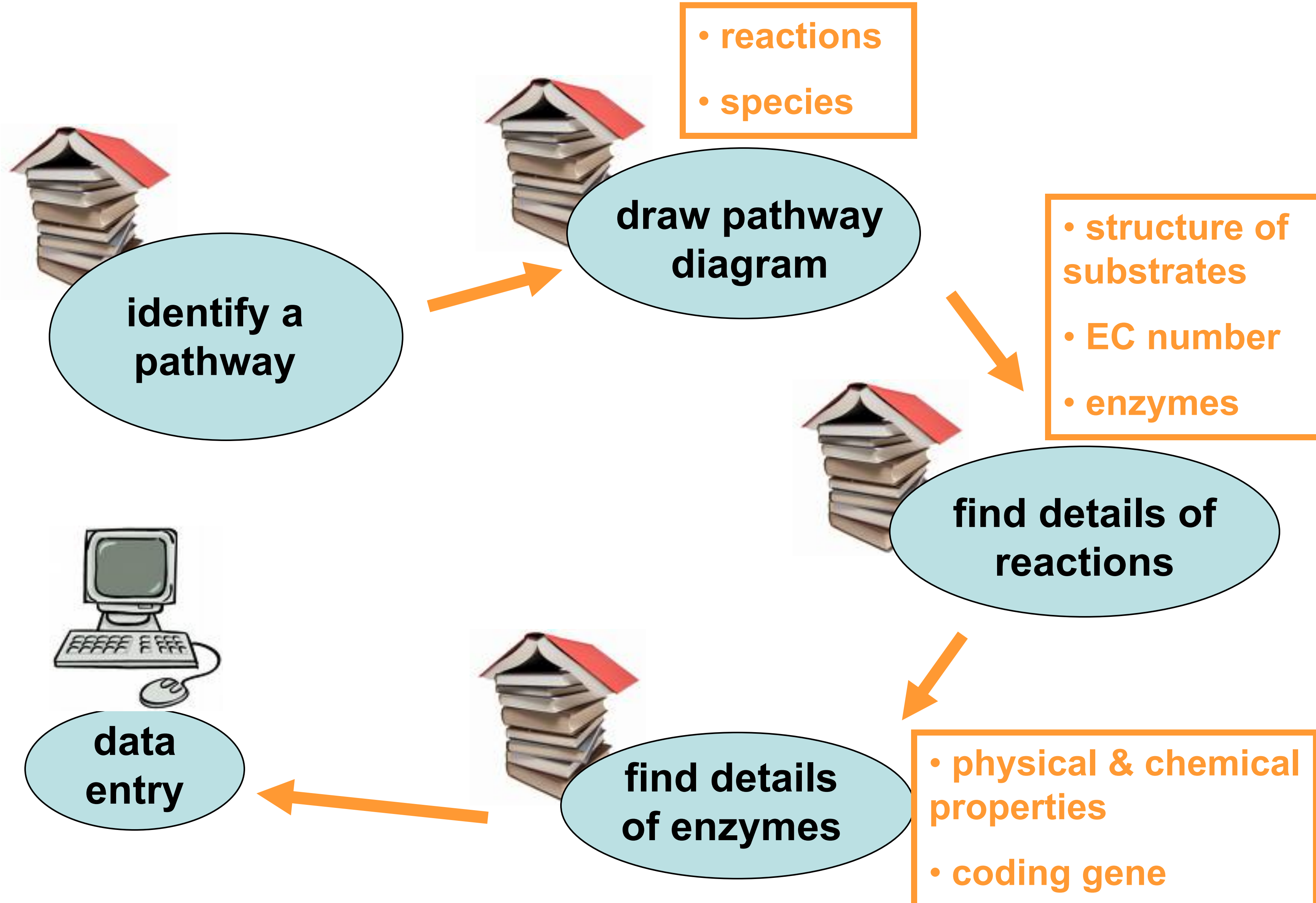
Curating missing pathways

- 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



identify a pathway

draw pathway diagram

- reactions
- species

find details of reactions

- structure of substrates
- EC number
- enzymes

find details of enzymes

- physical & chemical properties
- coding gene

data entry

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.

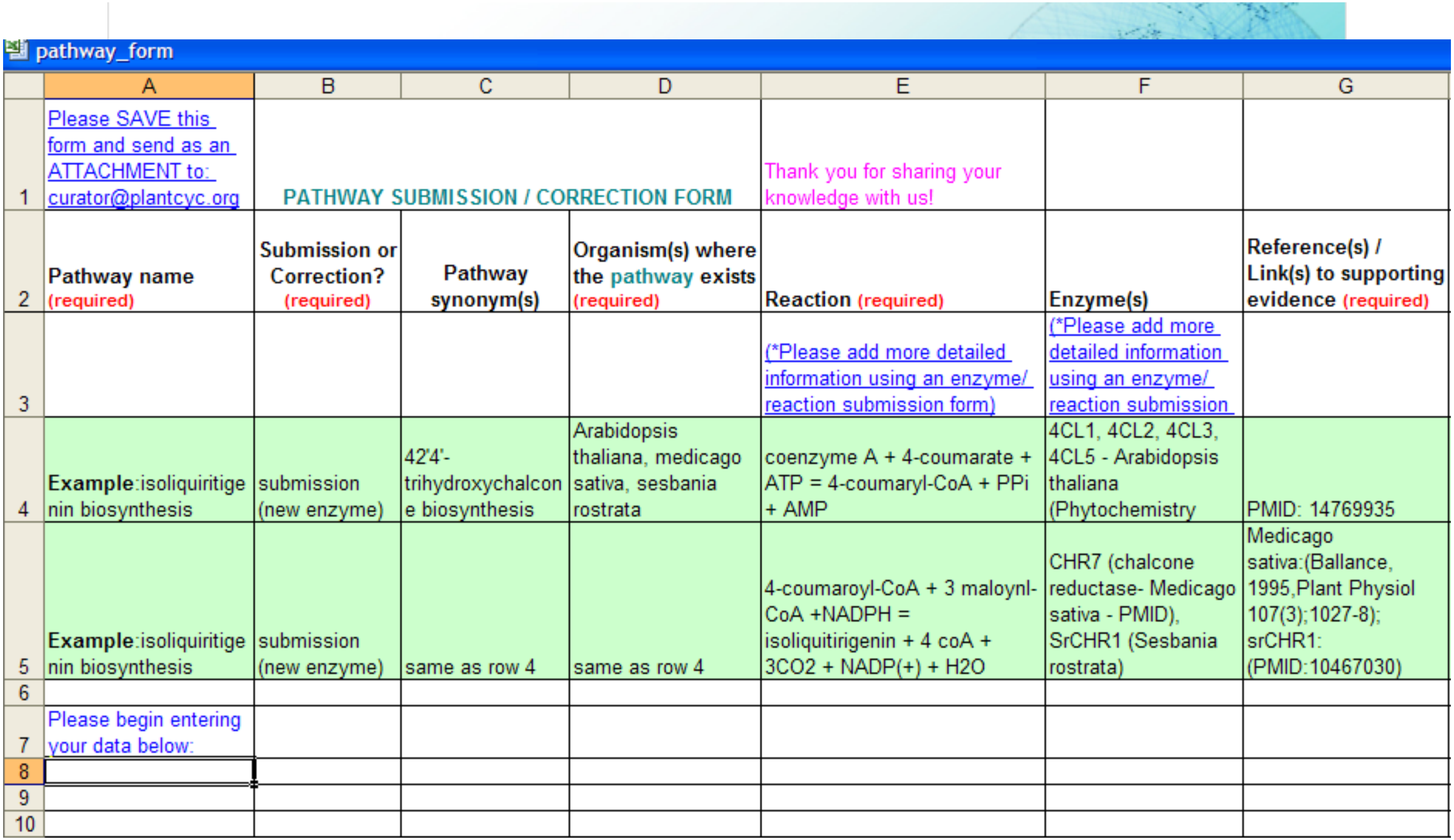
How you can help

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

- Enhance data accuracy
  - Reporting errors

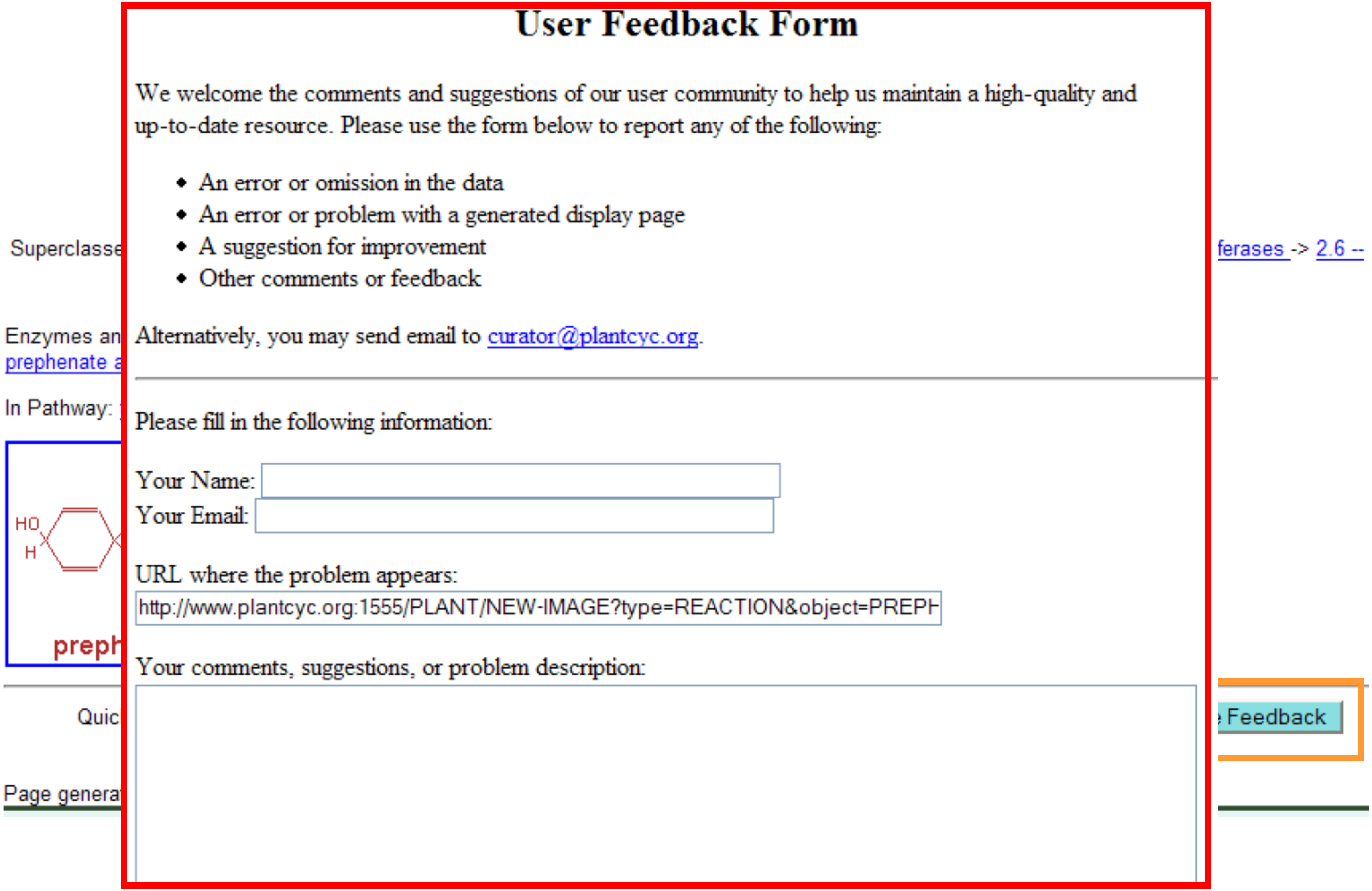
- Your idea/need of new features and functionalities

Data submission forms



pathway_form

	A	B	C	D	E	F	G
1	Please SAVE this form and send as an ATTACHMENT to: curator@plantcyc.org	PATHWAY SUBMISSION / CORRECTION FORM			Thank you for sharing your knowledge with us!		
2	**Pathway name** (required)	**Submission or Correction?** (required)	**Pathway synonym(s)**	**Organism(s) where the pathway exists** (required)	**Reaction** (required)	**Enzyme(s)**	**Reference(s) / Link(s) to supporting evidence** (required)
3					(*Please add more detailed information using an enzyme/ reaction submission form)	(*Please add more detailed information using an enzyme/ reaction submission	
4	**Example**:isoliquiritigenin biosynthesis	submission (new enzyme)	42'4'-trihydroxychalcone biosynthesis	Arabidopsis thaliana, medicago sativa, sesbania rostrata	coenzyme A + 4-coumarate + ATP = 4-coumaryl-CoA + PPi + AMP	4CL1, 4CL2, 4CL3, 4CL5 - Arabidopsis thaliana (Phytochemistry	PMID: 14769935
5	**Example**:isoliquiritigenin biosynthesis	submission (new enzyme)	same as row 4	same as row 4	4-coumaroyl-CoA + 3 maloynl-CoA +NADPH = isoliquitirigenin + 4 coA + 3CO2 + NADP(+) + H2O	CHR7 (chalcone reductase- Medicago sativa - PMID), SrCHR1 (Sesbania rostrata)	Medicago sativa:(Ballance, 1995,Plant Physiol 107(3);1027-8); srCHR1: (PMID:10467030)
6							
7	Please begin entering your data below:						
8							
9							
10							



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:

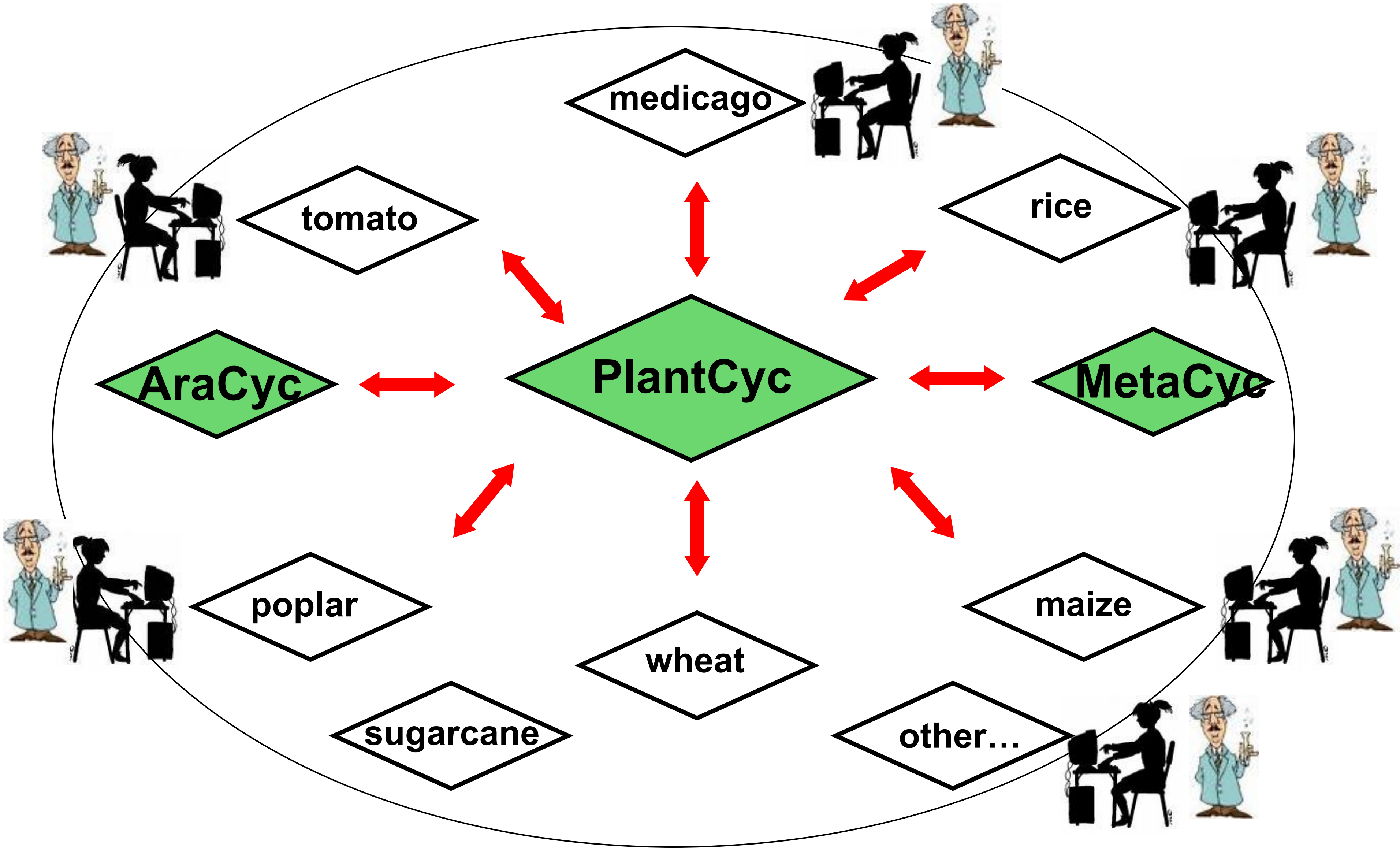
http://www.plantcyc.org:1555/PLANT/NEW-IMAGE?type=REACTION&object=PREPH

Your comments, suggestions, or problem description:

Email to us

- curator@plantcyc.org

The PMN project, us and you



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