Editing Pathway/Genome Databases I

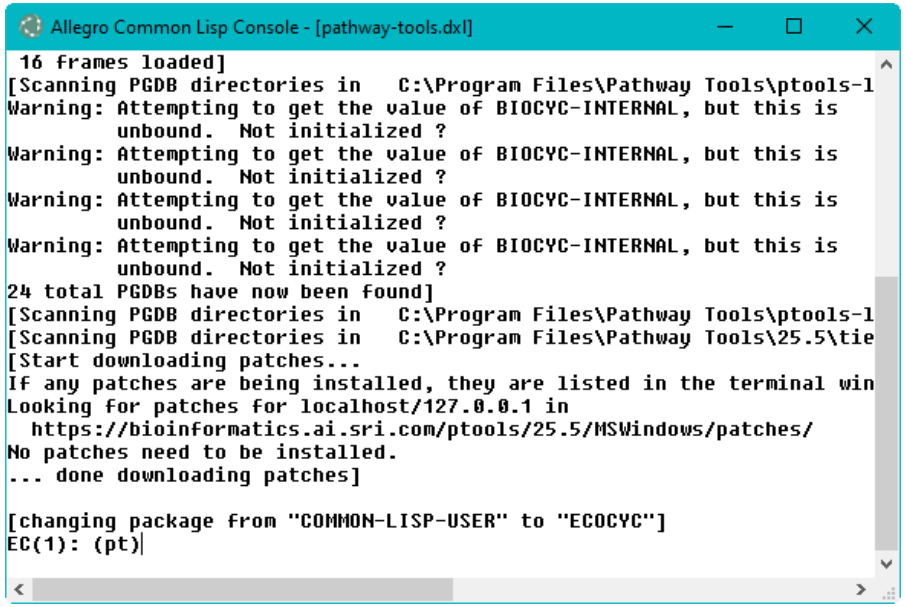
By Ron Caspi

ron.caspi@sri.com

A copy of this presentation could be found at
http://bioinformatics.ai.sri.com/ptools/tutorial/sessions/curation

Starting Pathway Tools

- Start the software with the Lisp option
- The Lisp console will show up
- At the prompt type (pt) (followed by Enter) to start the graphical user interface (GUI)
- If you close the GUI, the Lisp console will continue to run
- Exit by typing (exit)

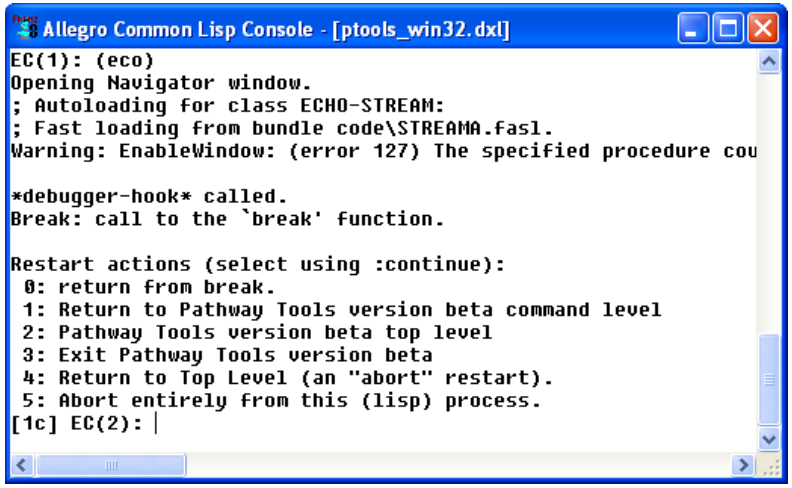


Working in a Lisp environment: breaks

- A break shifts the focus from the main GUI to the Lisp console
- To generate a break: BREAK key on your keyboard. If you don't have one, type (break) at the listener pane and hit enter
- When Lisp encounters an unrecognized command, it breaks
- A break is NOT a crash

Lisp presents several recovery options from a break

Type :cont x where x is the number of the best option



Bug reports

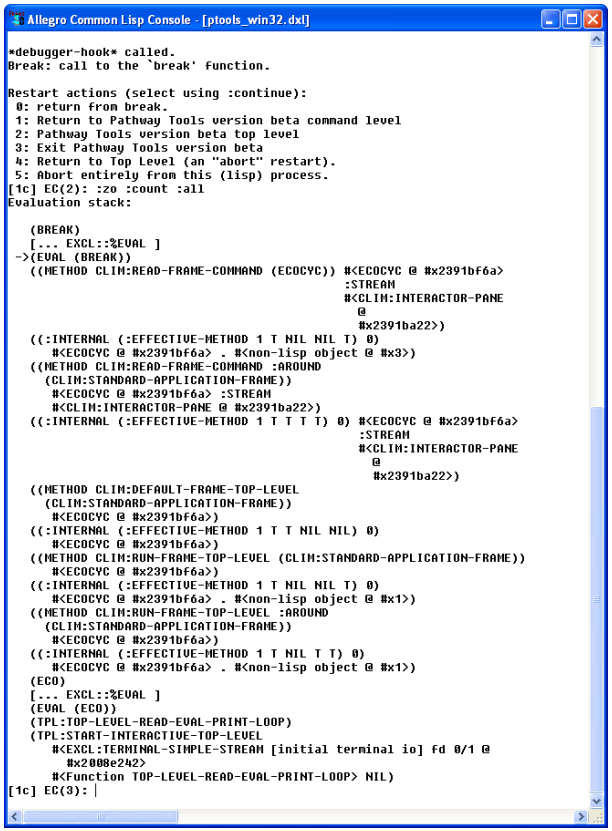
If you get a break as a result of a bug, get the evaluation stack by typing

:zo :count :15 at the lisp prompt

Copy the output, and send it

by email to

ptools-support@ai.sri.com

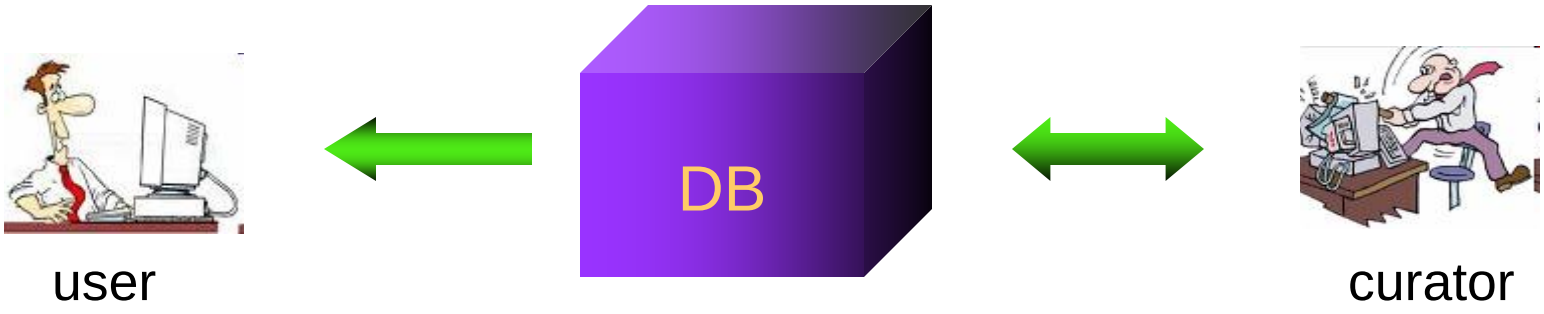


Pathway Tools in editing mode

The database can be accessed by two distinct modes

- Navigator mode allows no modification of the DB
- Editing mode allows complete modification of the DB

Editing is currently available in Desktop mode only, though online editors are being developed



Installing an editable PGDB

In order to be able to perform editing, you must have an editable PGDB installed on your system (built-in PGDBs can't be edited).

In the following exercises we will be using the PGDB for *Arthrospira platensis NIES-39*.

Your installation has this PGDB built-in. To make an editable copy of it:

- Select this PGDB
- File → Save PGDB As…
- Type "Test" for New PGDB ID and click outside that box.
- Click OK and wait for the process to complete.
- You will now have two PGDBs for this organism. Open the one with "Source" listed as "User".

Pathway Tools -- Available Databases

Organisms	Pathways	Genes (ORF %)	Genome Size (bp)	Citations	Source	Version	Registry Download Date
Arabidopsis thaliana col					User (MySQL)	24.0	
Arthrospira platensis NIES-39	228	6,579 (38.4%)	6,788,435	1759	User	25.5	
Arthrospira platensis NIES-39	228	6,579 (38.4%)	6,788,435	1759	Built-In	25.5	
Escherichia coli K-12 substr. MG1655	363	4,735 (10.9%)	4,641,652	41490	Built-In	25.5	
Helicobacter pylori 26695					Registry	25.5	23-Feb-2022 14:49:12
MetaCyc	2,980	14,343 (0.5%)	0	70088	Built-In	25.5	

Copyright Notice

Saving/undoing changes

The user **must** save changes explicitly

- File → Save Current DB

  or
- Save DB button on upper right corner

  or
- ^s

"Undo" is called Revert Current DB in PTools lingo. It only works with unsaved changes, and it reverts **all** unsaved changes (no step-by step undo).

Storing databases in MySQL or Oracle enables the following commands:

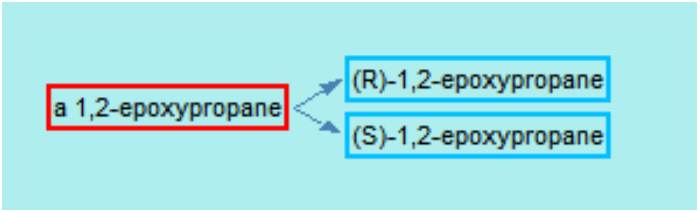
- List Unsaved Changes in Current DB
- Checkpoint Current DB Updates to File
- Restore Updates from Checkpoint File
- Refresh All Open DBs

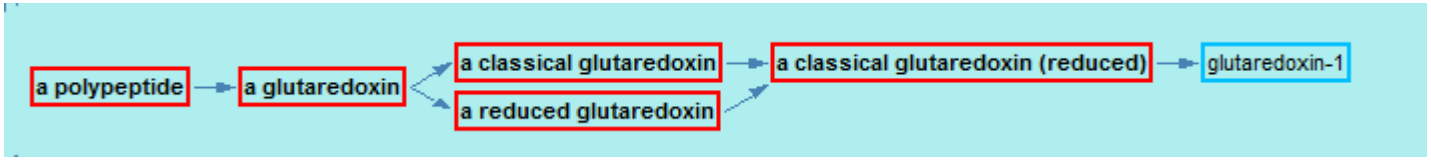
Other editing-related DB commands under the File menu

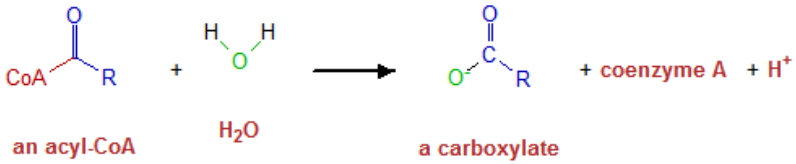
- Create new version for selected DBs (also modifies the default-version file to have the new version opened automatically)
- Save DB as… (makes a new copy that can be opened in the same session as the source PGDB)
- Delete a DB

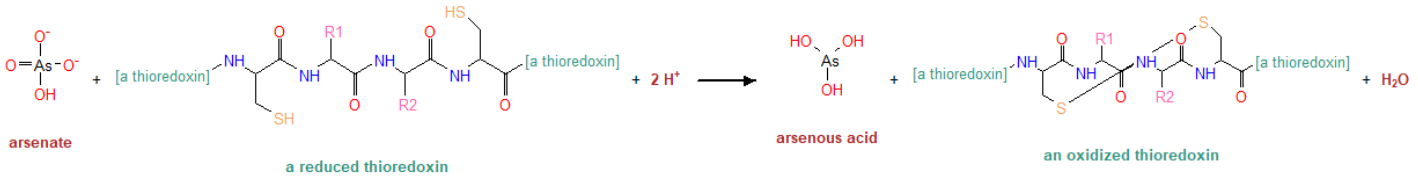
Classes and instances are like folders and files

- Instances describe specific objects (e.g. L-lysine)
- Classes describe groups of biological objects (e.g. “an amino acid”)
- Classes can contain other objects, while instances can’t
- Every compound with an “R” in its structure should be a class
- Proteins or modified proteins that are substrates of MetaCyc reactions are always classes









What's a frame? And why does it need an ID?

- Every object in the database is a "frame", and each frame has a unique ID within the database. Instance frame IDs are usually assigned automatically and are not very meaningful.

  Examples: CPD-23 PWYQT-7 RXN0-555 MONOMER-387 CPLXI-345.

  The prefix describes the type of object. Frame names generated in a PGDB other than MetaCyc include one or two characters that identify the source database.

  Legacy objects in Pathway Tools (created before current naming standards) usually deviate from these guidelines.

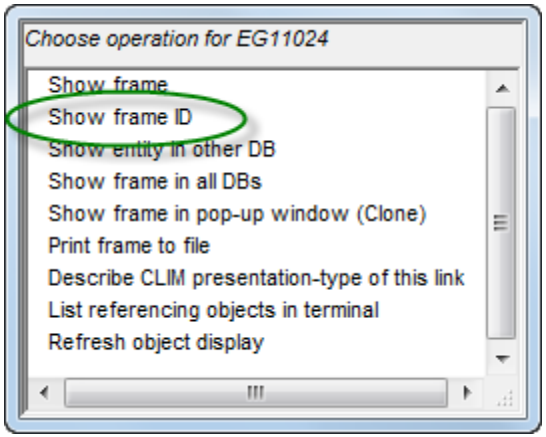
- Class frame names are usually created by humans and use language. E.g. Thioglucosides, Amino-Acid-Biosynthesis.

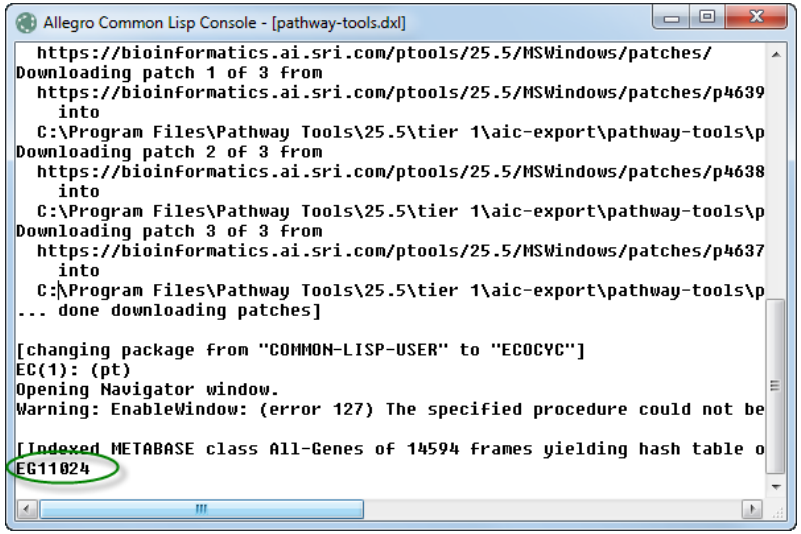
- Object **names** include common names and synonyms. They are useful for both humans and computer searches. Unlike frame IDs, names may be not unique.

Printing frame IDs in the Lisp Console

Curation involves working with frame IDs. A convenient way to do it is to print them to the Lisp console and copy them from there.

- Right-click on an object name and select Show → Show frame name
- Move to the Lisp console and copy the name to the clipboard.
- If collecting multiple names, paste it into a text editor



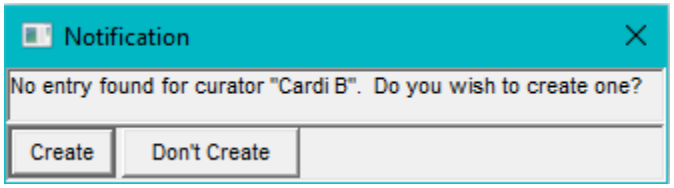


The author credit system

Author credit gives you credit for your work.

Associating objects with their author makes it easy to find the objects you worked on.

Author credit is not properly stored unless a curator frame has been created and configured properly.



***Helicobacter pylori* 26695 Curator: Ron Caspi**

Email: caspi@ai.sri.com

Affiliations: **SRI International**

Note: Listed below are contributions the curator has made to HpyCyc. They are sorted, with the most recent at the top.

Stats: Pathways: 48, Proteins: 536, RNAs: 59, Reactions: 134, Compounds: 1584, Misc.: 19

Pathways

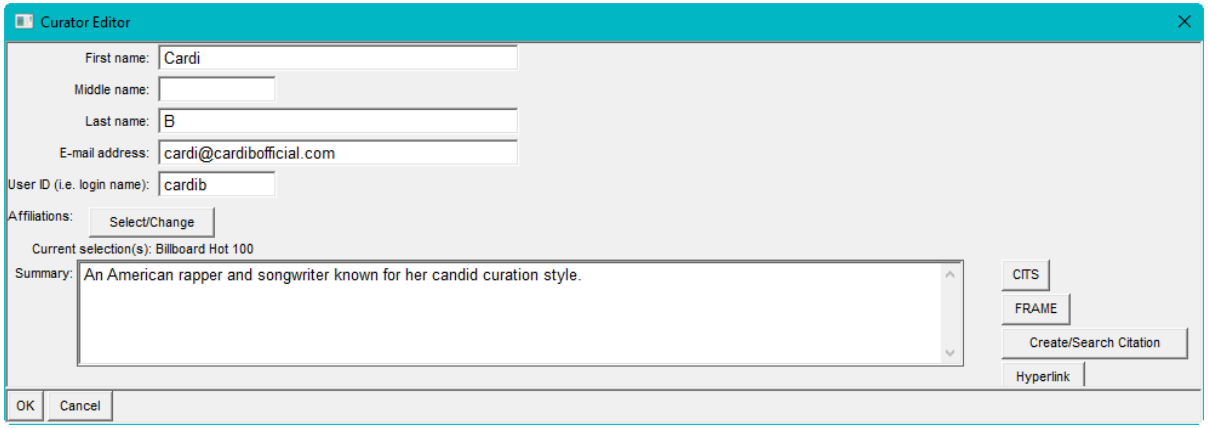
Shown below are the following credit types: Created: 25, Revised: 23

Created:

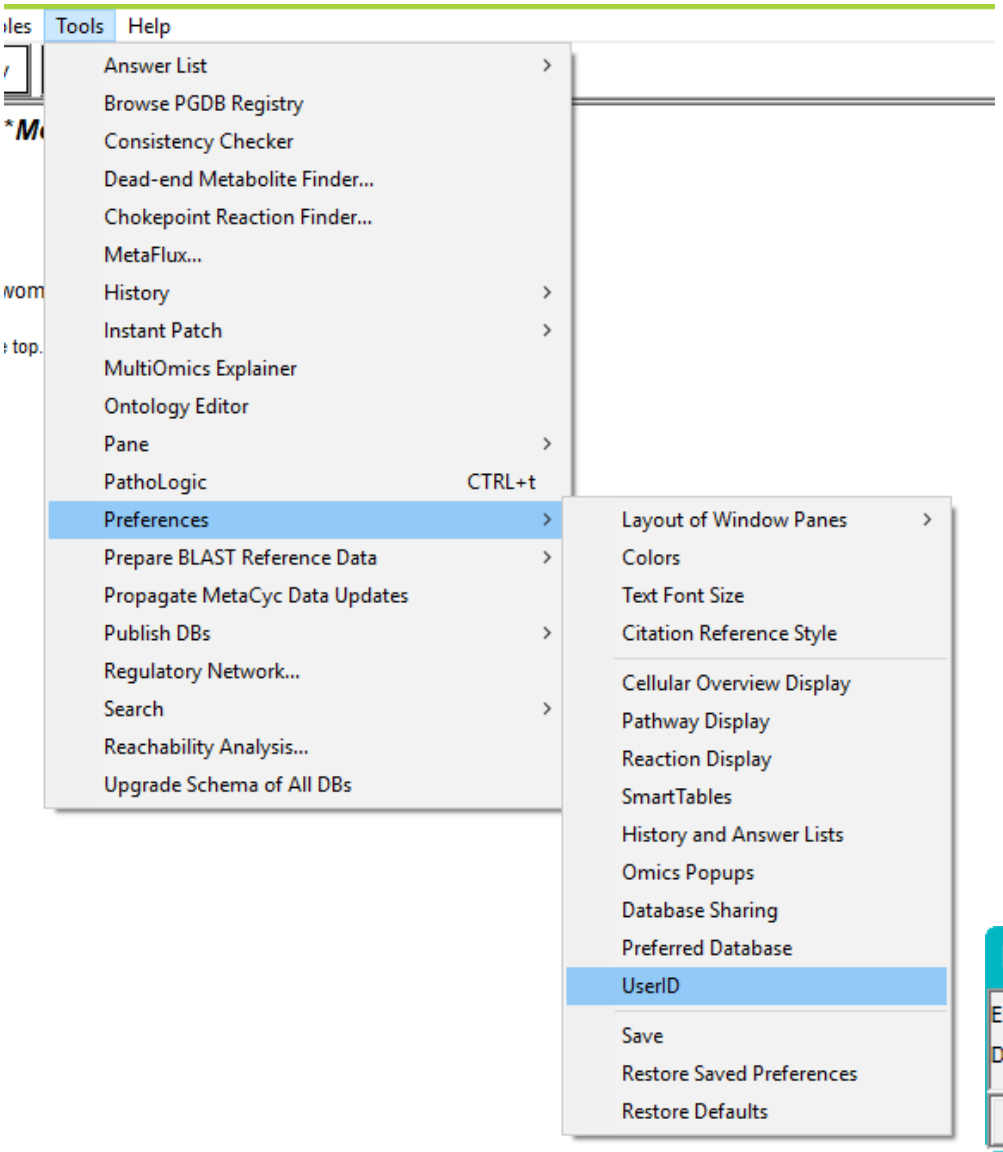
ATP biosynthesis on 19-Dec-2017,
formaldehyde oxidation VII (THF pathway) on 19-Jun-2017,
lipoprotein posttranslational modification on 10-Apr-2017,
(5Z)-dodec-5-enoate biosynthesis II on 24-Jan-2017,
biosynthesis of Lewis epitopes (H. pylori) on 21-Sep-2016,
L-cysteine biosynthesis VI (from L-methionine) on 30-Mar-2016,
crotonate fermentation (to acetate and cyclohexane carboxylate) on 02-Nov-2013,
superpathway of demethylmenaquinol-6 biosynthesis II on 17-Sep-2013,
demethylmenaquinol-6 biosynthesis II on 17-Sep-2013,
superpathway of adenosine nucleotides *de novo* biosynthesis I on 18-Feb-2013,
superpathway of guanosine nucleotides *de novo* biosynthesis I on 15-Feb-2013,
superpathway of pyrimidine deoxyribonucleotides *de novo* biosynthesis on 31-Jan-2013,
pyrimidine deoxyribonucleotides *de novo* biosynthesis I on 16-Jan-2013,
2'-deoxy-α-D-ribose 1-phosphate degradation on 08-Jan-2013,
purine deoxyribonucleosides degradation I on 08-Jan-2013,
hydroxymethylpyrimidine salvage on 23-Sep-2011,
thiamine diphosphate salvage II on 14-Sep-2011,
tetrahydrofolate biosynthesis on 22-Sep-2010,
autoinducer AI-2 biosynthesis I on 09-Mar-2009,
CMP-pseudaminate biosynthesis on 25-Feb-2009,
superpathway of adenosine nucleotides *de novo* biosynthesis II on 13-Jan-2009,
superpathway of guanosine nucleotides *de novo* biosynthesis II on 13-Jan-2009,
acyl carrier protein activation on 13-Aug-2008,
UMP biosynthesis I on 24-Oct-2007,
GDP-mannose biosynthesis on 13-Sep-2007

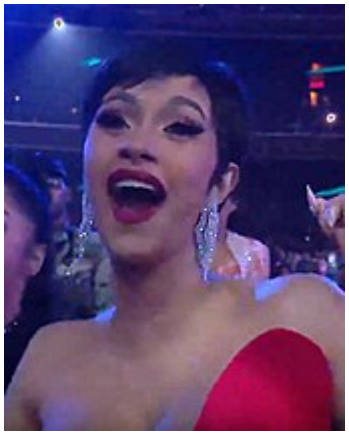
Creating curator frame – part 1

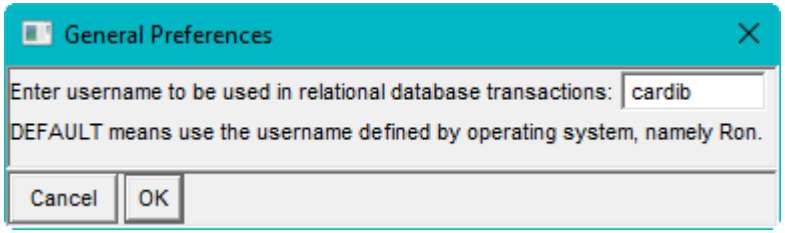
- Start by checking if there already is a frame for your organization (Tools → Search → Organizations)
- If there isn't one, create one (File → Create → Organization).
- Next, create a frame for yourself (File → Create → Curator)
- User ID must be longer than 3 characters



Creating curator frame – part 2







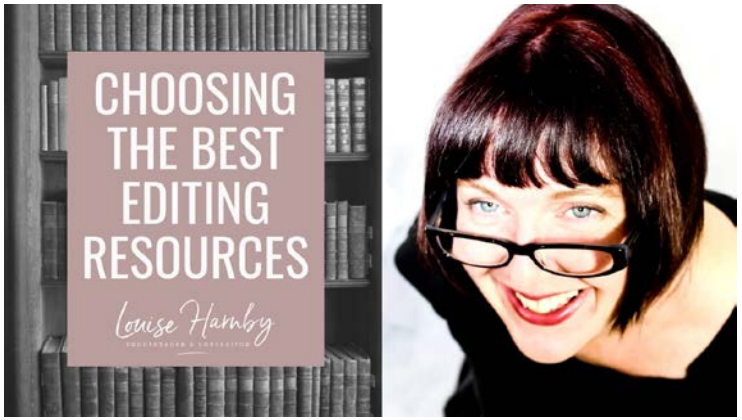
What sort of objects are usually edited?

The following are just a few examples

- metabolites
- reactions
- pathways
- enzymes: (creating complexes, assigning to correct reactions)
- regulatory information
- transcription units

The editors

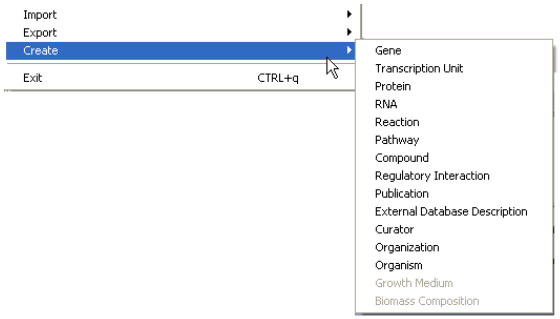
- PGDB Info Editor
- Compound Editor and Compound Structure Editors
- Reaction Editor
- Pathway Editor, Pathway Info Editor
- Signaling Pathway Editor
- Protein Editor and Protein Subunit Structure Editor
- Synonym Editor
- Publication Editor
- Curator/Organization Editors
- Gene Editor
- Isoform/Coding-Segment Editor
- RNA Editor
- Transcription Unit Editor
- Regulatory Interaction Editor
- External Database Editor
- Organism Editor
- Frame Editor
- Ontology Editor



Invoking the Editors

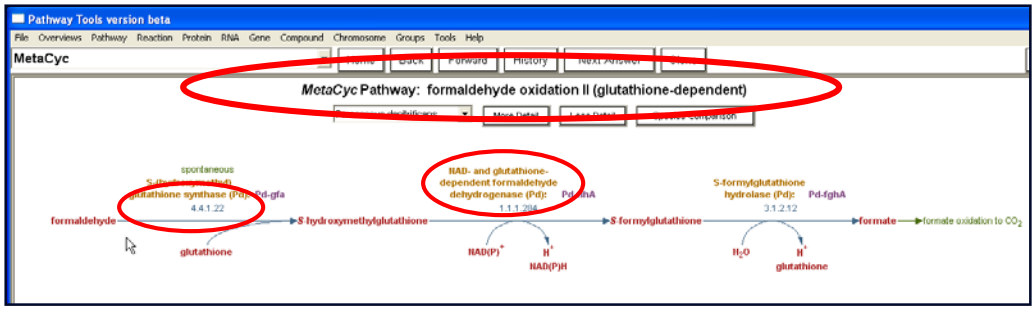
Creating a new Object:

Use the New command under certain top menus, or the Create command under the File menu

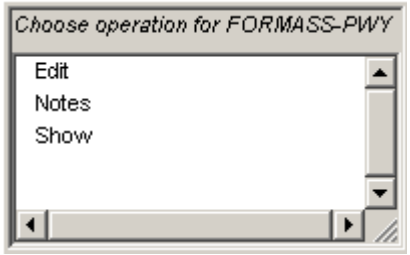


Editing an existing Object:

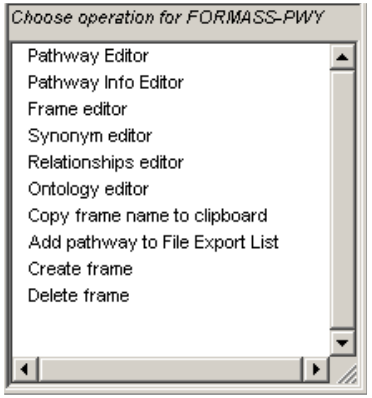
Right-Click on the any clickable name, select Edit, then the appropriate editor



Right-click



left-click on “Edit”

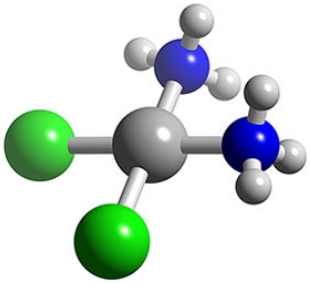


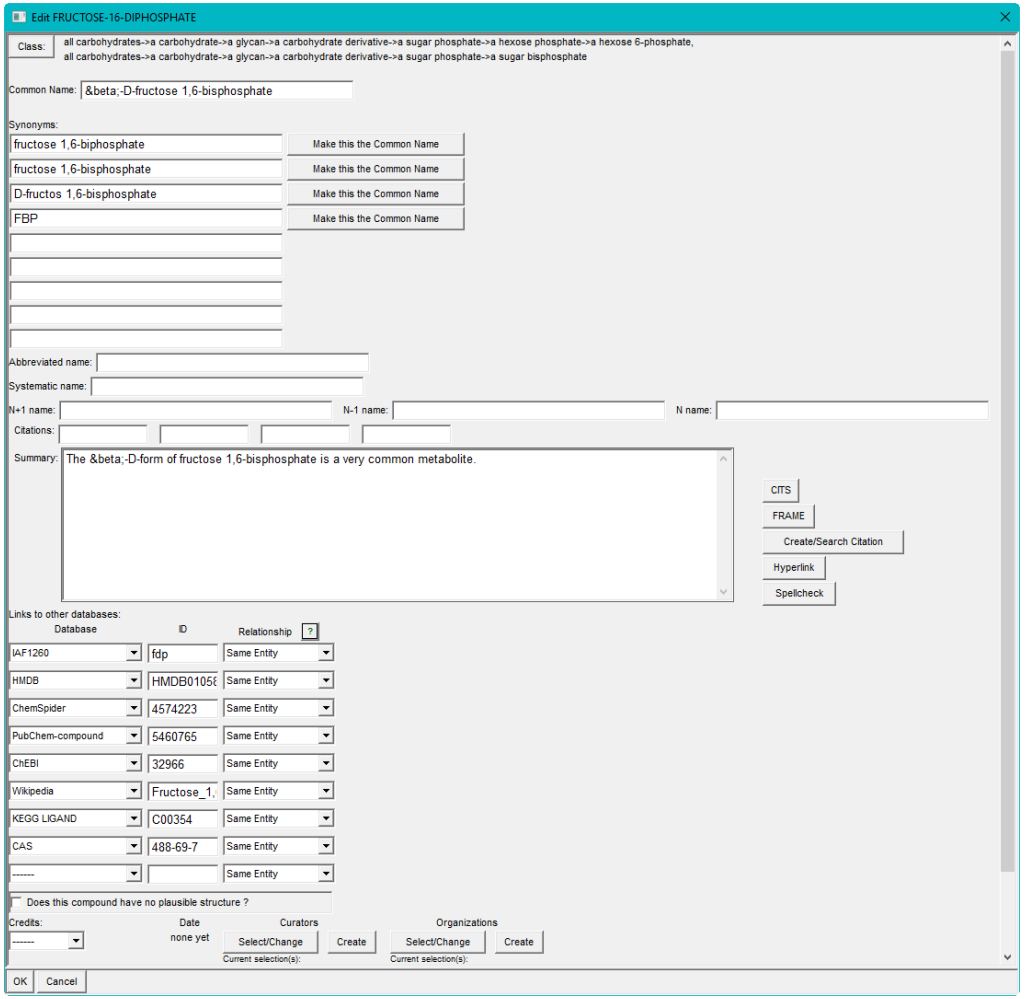
COMPOUNDS

The compound editor

- Create or edit a compound (but not its structure)

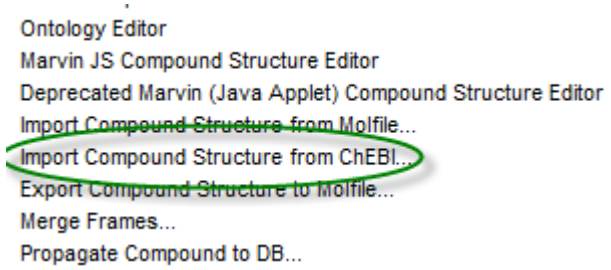
- Specify Class
- Common Name and Synonyms
- Comments, citations
- Links to other DBs



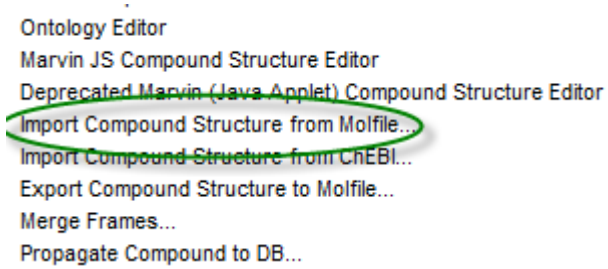


Entering compound structures

- Option 1: if the compound is present in ChEBI (https://www.ebi.ac.uk/chebi/) and you enter the link, you can import the structure directly from ChEBI



- Option 2: install a freeware desktop program such as ChemSketch, save structure as Mol file, and import into PTools



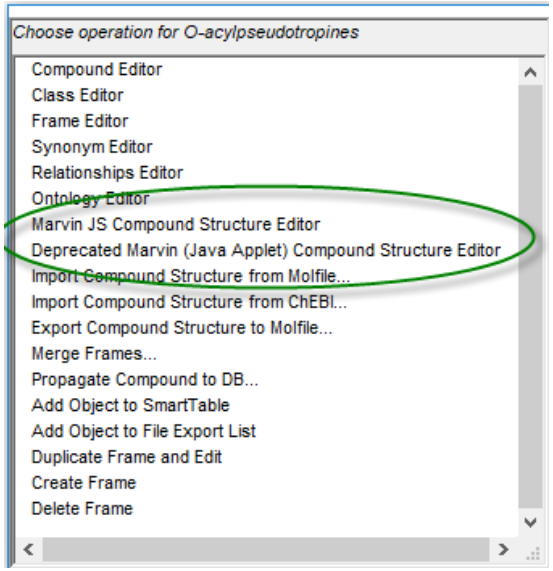
https://www.acdlabs.com/resources/freeware/chemsketch/

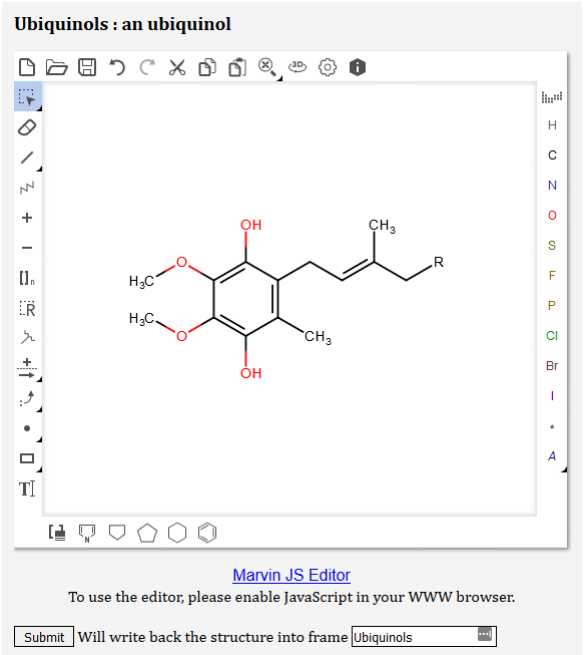
The Marvin structure editor

- Pathway Tools supports the Marvin JS structure editor, produced by ChemAxon, which needs to be obtained from them.

https://chemaxon.com/products/marvin-js

http://bioinformatics.ai.sri.com/ptools/installation-guide/released/marvin-js.html





Other compound-related functions

- Exporting to other DBs
- Duplicate Frame and Edit

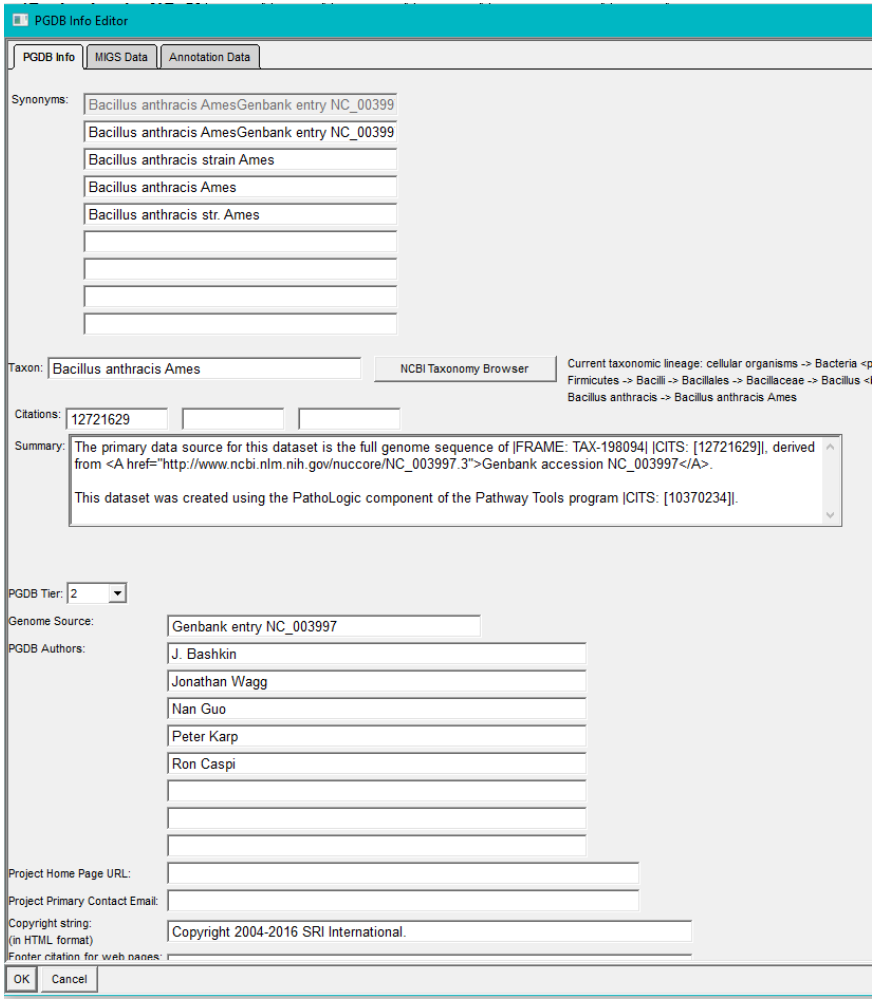
The PGDB info editor

To access: go to the PGDB home page (File-> Summarize Current Database or Organism Set)

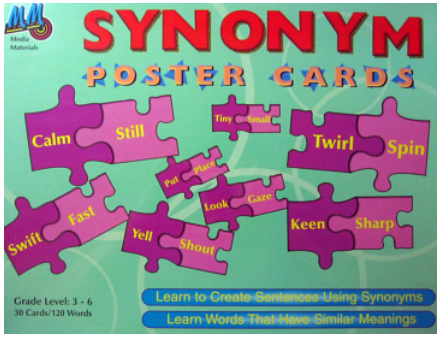
Right-click the organism name, and choose Edit → PGDB Info Editor

This is the place to:

- Create a comment for the PGDB home page
- Specify PGDB authors
- Modify NCBI taxonomy
- Specify a footer
- Set the tier level
- Enter MIGS Data
- Enter Annotation Data

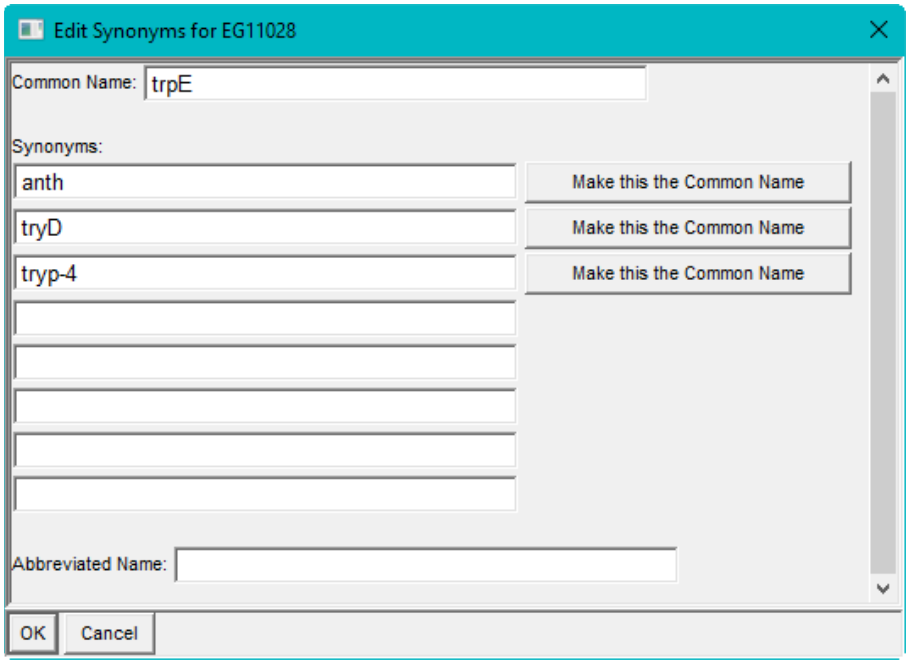


The synonym editor



Lets you easily edit the synonyms and set the common name

The more synonyms, the more likely users are to find an object



Correct assignment of reactions to enzymes

After running PathoLogic, some enzymes will be assigned to reactions incorrectly. A curator needs to remove incorrect assignments and attach correct ones.

Example: gene NIES39_RS19045 (*ubiE*) was curated by RefSeq as bifunctional demethylmenaquinone methyltransferase/2-methoxy-6-polyprenyl-1,4-benzoquinol methylase UbiE, and thus assigned multiple reactions of EC 2.1.1.163, demethylmenaquinone methyltransferase.

Since cyanobacteria produce plastoquinone and not menaquinone, this can't be true.

Click on the UniProt link and you will find that in fact this is MenG and should be assigned as EC 2.1.1.329 (a BLAST search against any cyanobacterium would have produced this result as well). This reaction is part of the phylloquinol biosynthesis pathway.

UbiE or MenG?

UbiE/MenG has been assigned 5 incorrect reactions. Each one is involved in an incorrectly-predicted pathway. We will

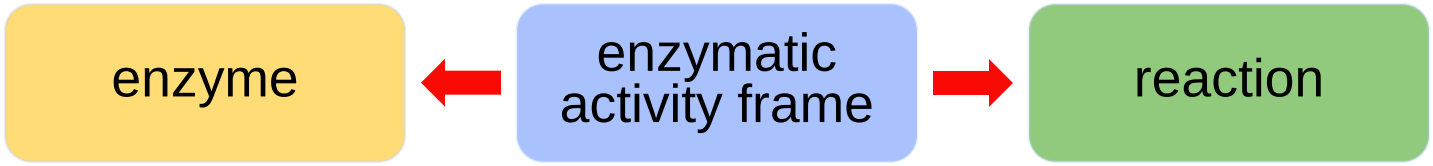
1. Change the gene name to menG
2. Change the enzyme name to demethylphylloquinol methyltransferase
3. Delete the incorrect pathways
4. Delete the incorrect reactions
5. Attach the enzyme to the correct reaction

Other things to do here (but we won't today)

1. Import the plastoquinone biosynthesis pathway from MetaCyc
2. Find enzymes known to participate in that pathway and attach them to the appropriate reactions

How enzymatic activities are handled

Each enzymatic activity is defined by a new database object that points to an enzyme and a reaction

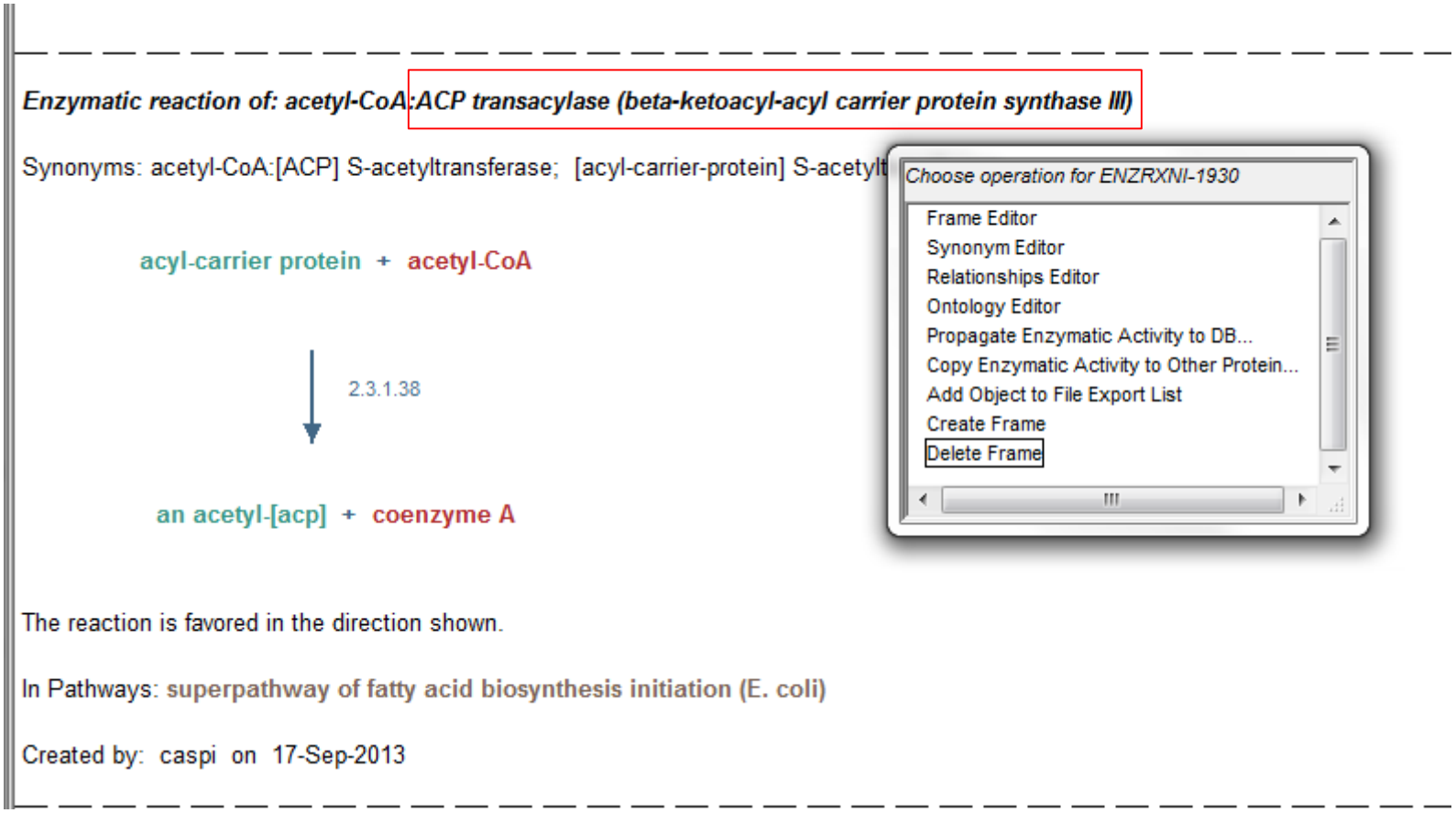


The enzymatic activity object is a frame, like everything else in the database. It's frame ID is in the form ENZRXN-XXX

Enzymatic activities also have a common name and potential synonyms (e.g. 2-oxoglutarate synthase)

Deleting incorrect enzymatic activities

Right-click the enzymatic activity name, and select Edit -> Delete



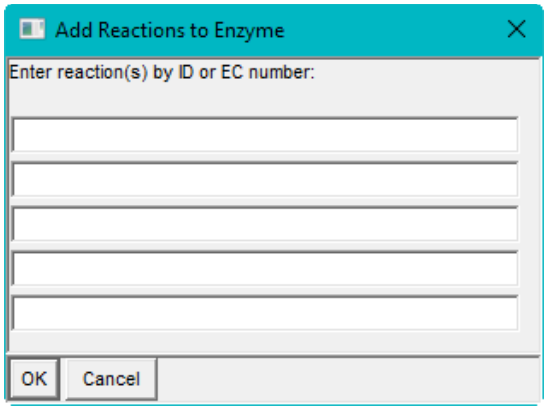
Attaching a reaction to an enzyme

To add an enzyme to a reaction: First copy the frame ID of the enzyme, then

- Right click the reaction, choose Edit → Create/Add enzyme and paste the ID.

Or

- Copy the frame ID of the reaction, then
- Right click the enzyme, choose Edit → Add Reaction(s) and paste the ID.

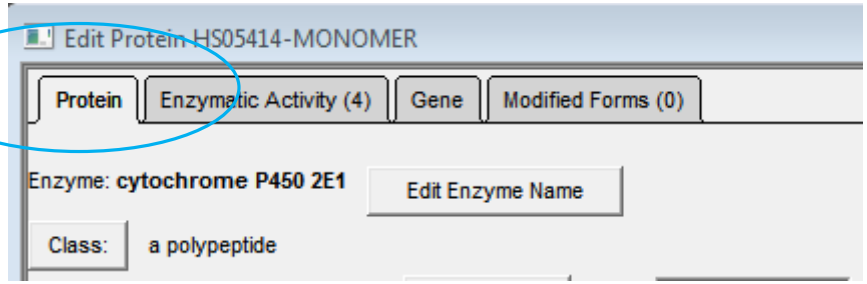


Connecting a reaction with a protein will open the Protein Editor and require you to enter an enzymatic activity name in the editor. More about that in the next slides.

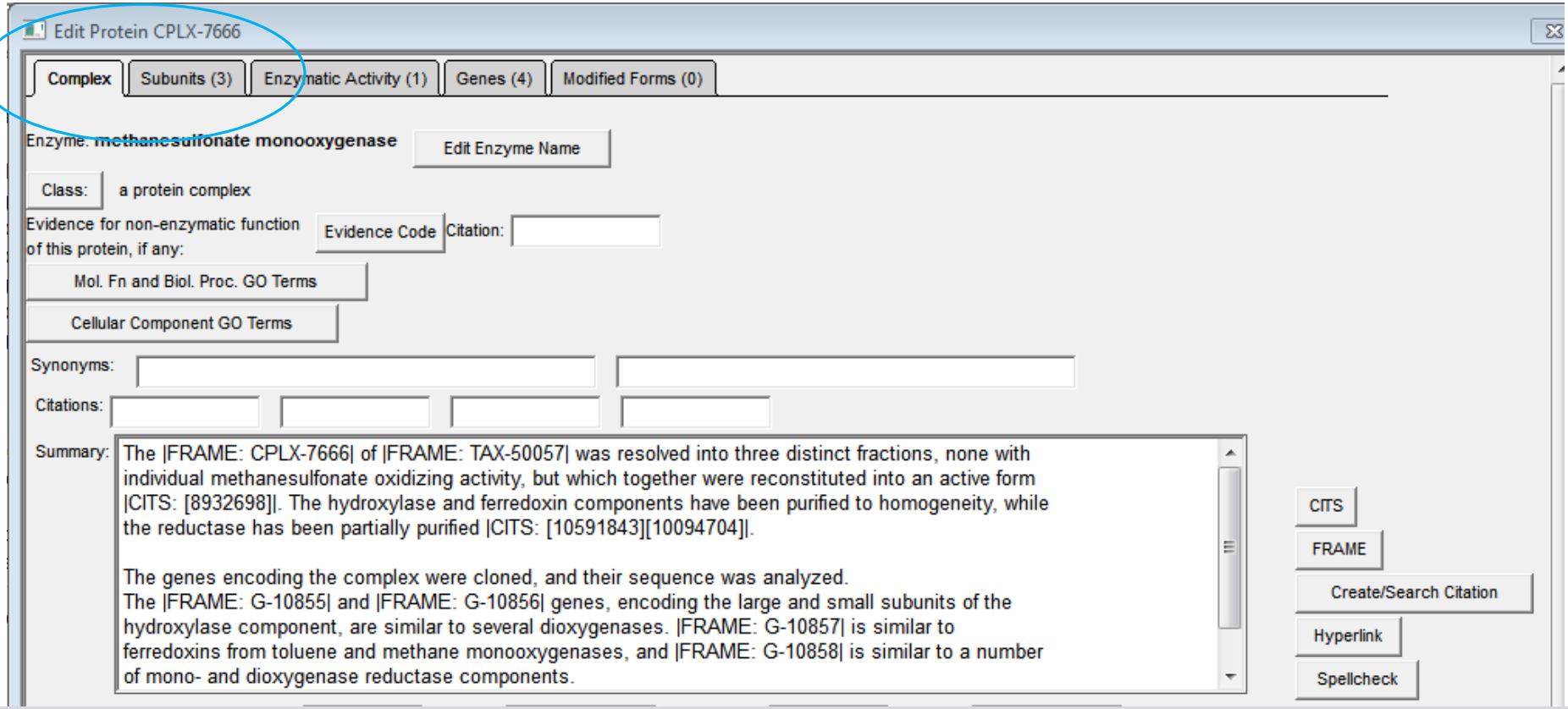
PROTEINS

The protein editor

(monomer version)

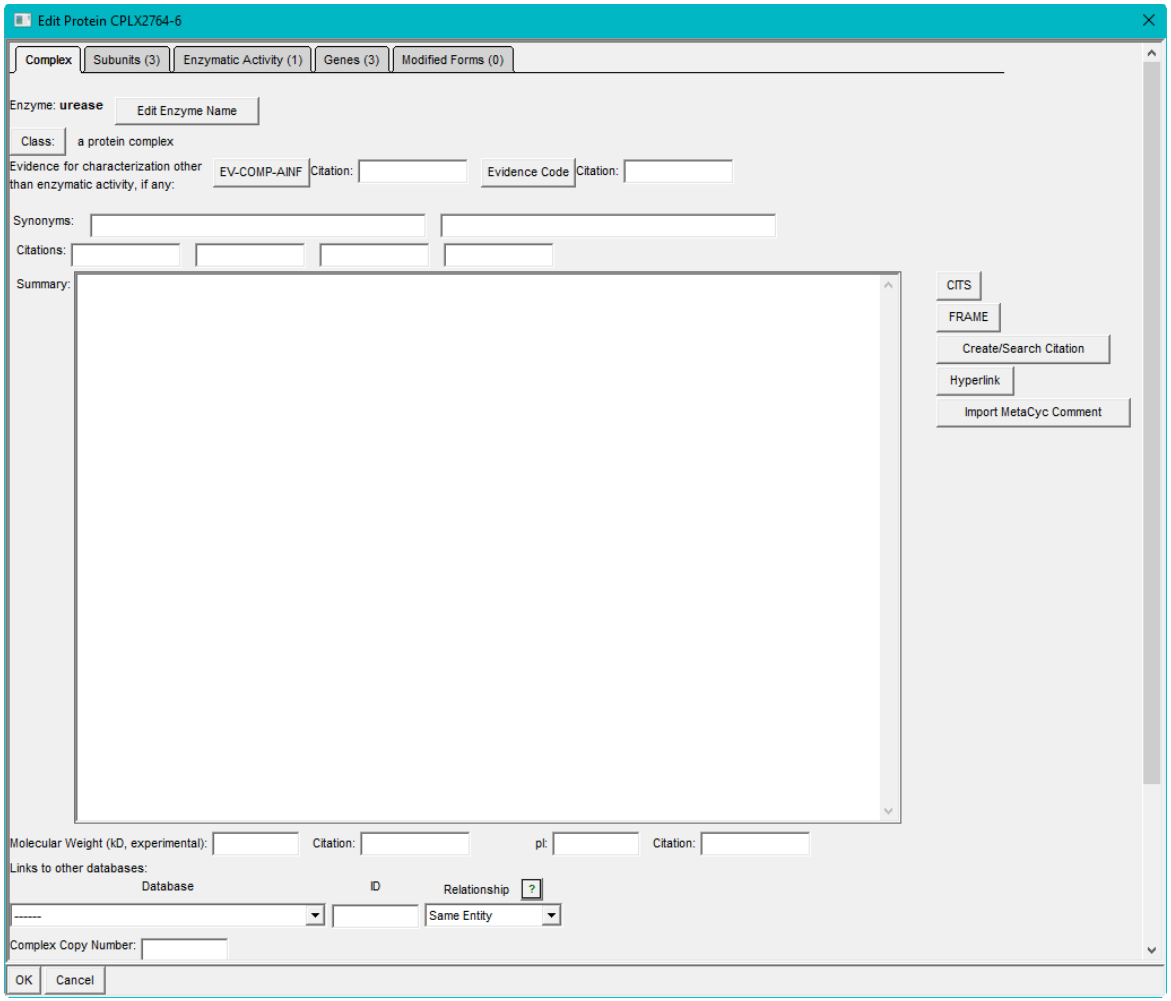


(protein complex version)



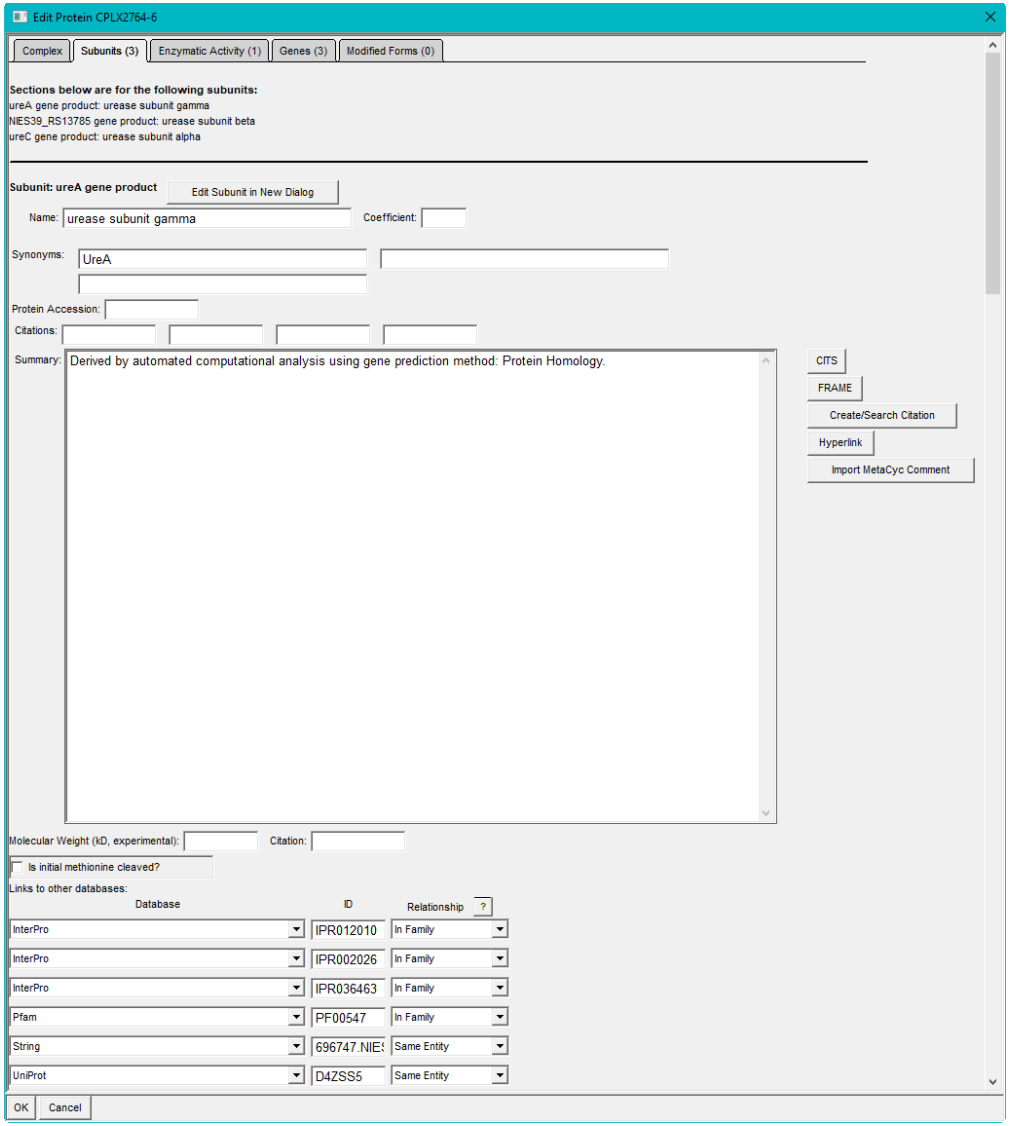
Protein editor – first tab

For an example of a complex, open CPLX2764-6 in the *A. platensis* PGDB

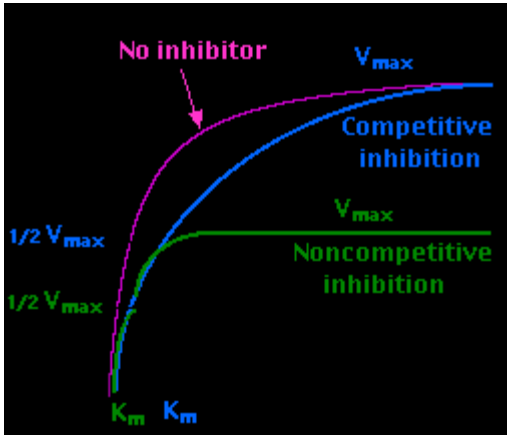


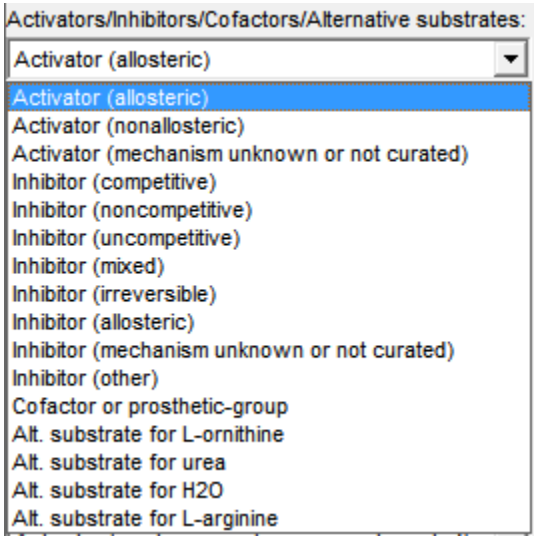
Protein editor - subunits tab (for complexes)

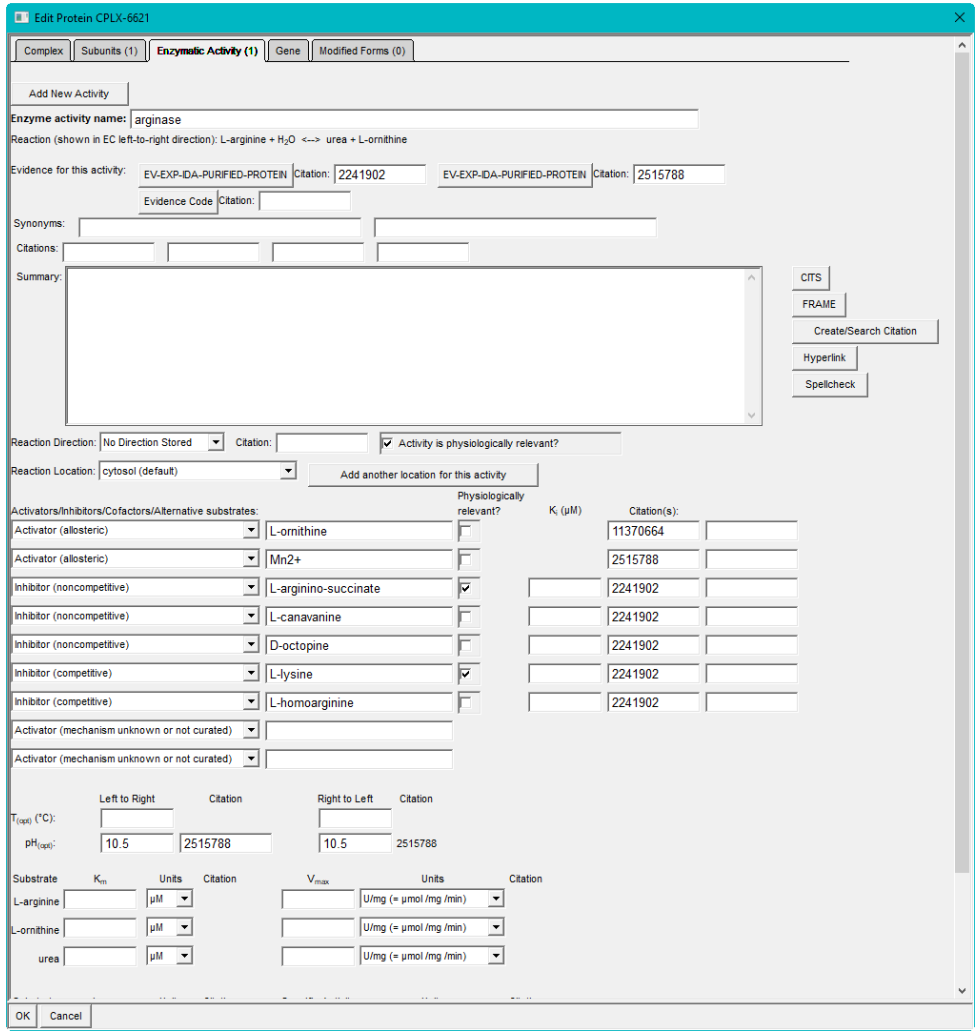
- Edit the copy number of each subunit
- Specify UniProt ID or links to other databases
- Specify experimental MW and any useful info that may apply (GO terms, features, copy number)



Protein editor - enzymatic activity tab

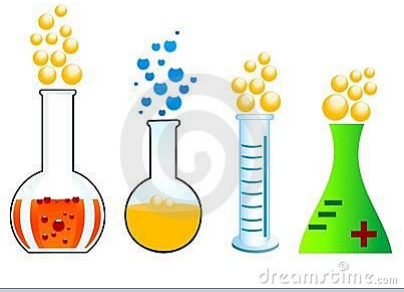






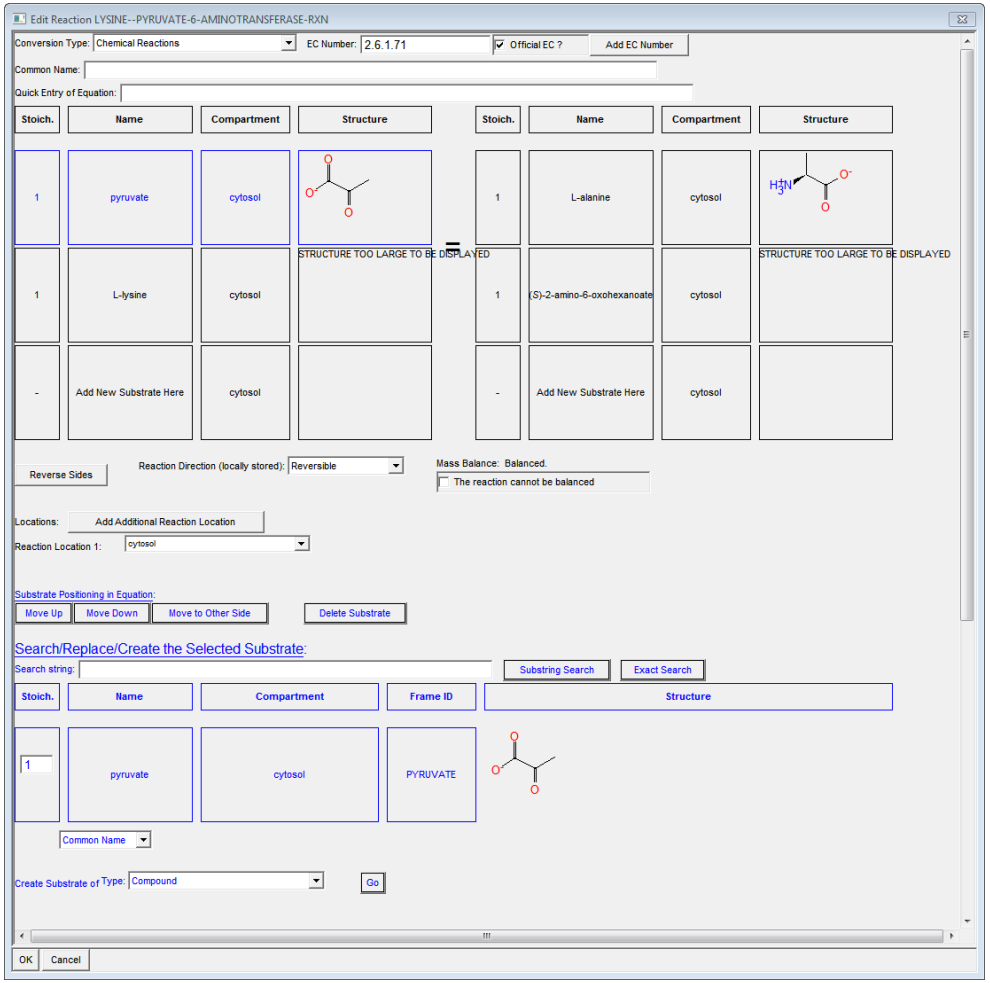
REACTIONS

The reaction editor



With the Reaction Editor you can:

- Enter or edit a reaction equation
- Specify EC numbers (official?)
- Enter a common name (if no full EC number exists)
- Set Conversion Type
- Specify location information (transport, cellular location)
- Specify reaction direction



L-lysine + H2O + O2 = 6-amino-2-oxohexanoate + ammonium + H2O2

Reaction editor examples

Regular chemical reaction:

L-lysine + H2O + O2 = 6-amino-2-oxohexanoate + ammonium + H2O2

Transport reaction:

first, set reaction type to transport

L-lysine[out] + ATP[in] + H2O[in] = L-lysine[in] + ADP[in] + Pi[in]
(note – unbalanced)

balance, then add reaction locations

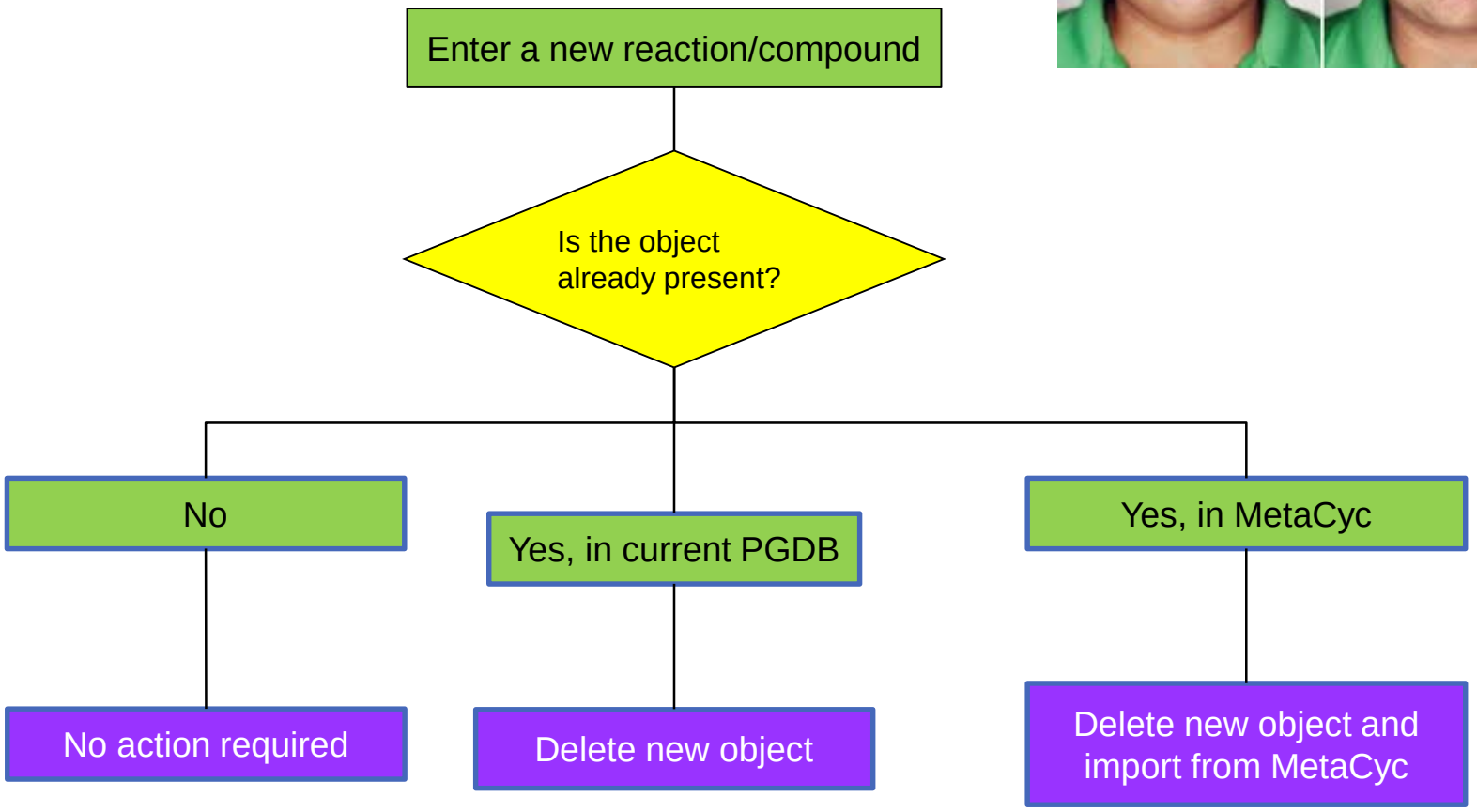
Duplication is bad

Avoid Duplication! Reuse information whenever possible

- A PGDB should not describe the same biological or chemical entity more than once
- You should not recreate an entry already present in MetaCyc
- Some tools help prevent the inadvertent creation of duplicate compounds and reactions

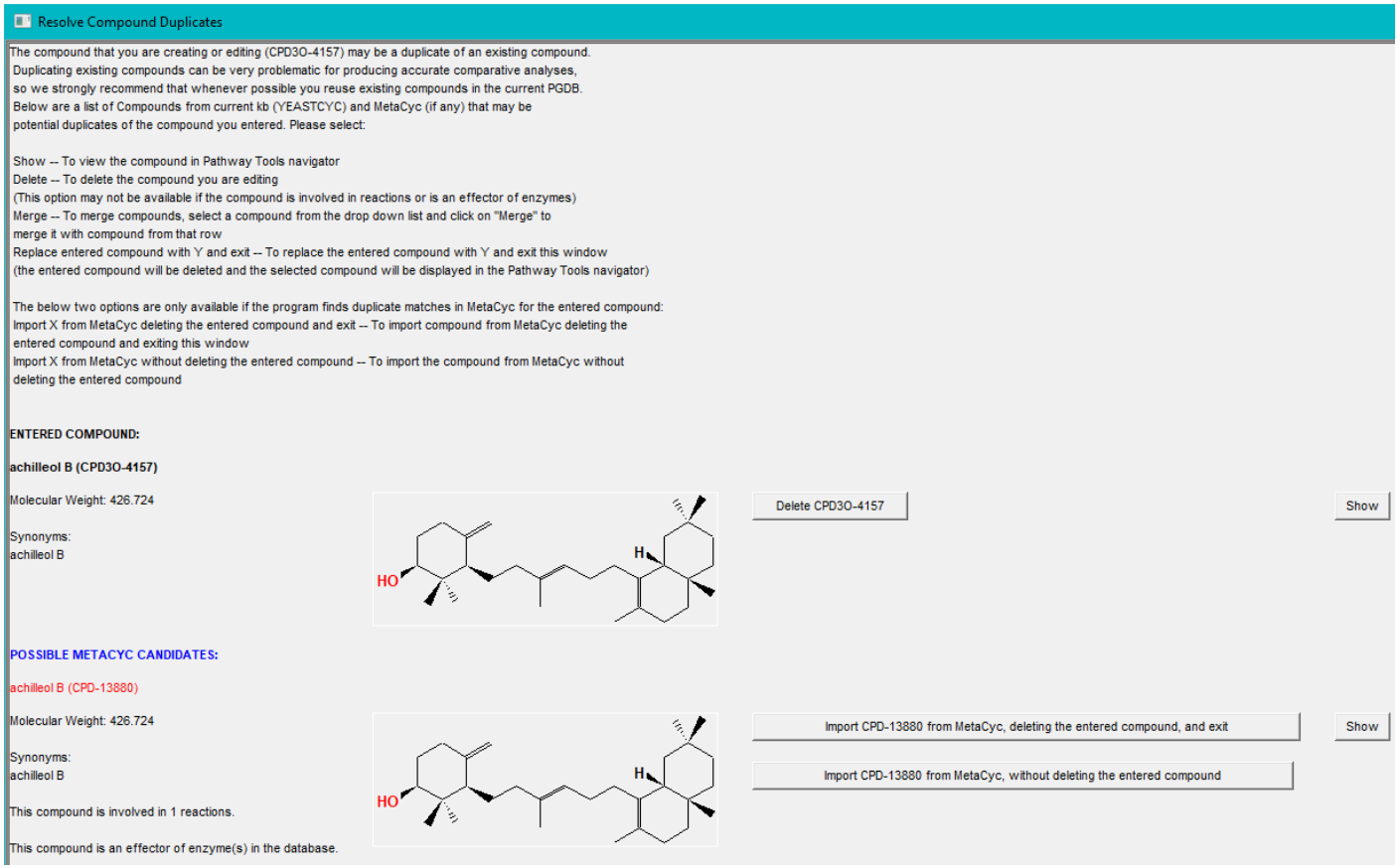
Duplicates can be cute, but should be avoided

The duplicate checker



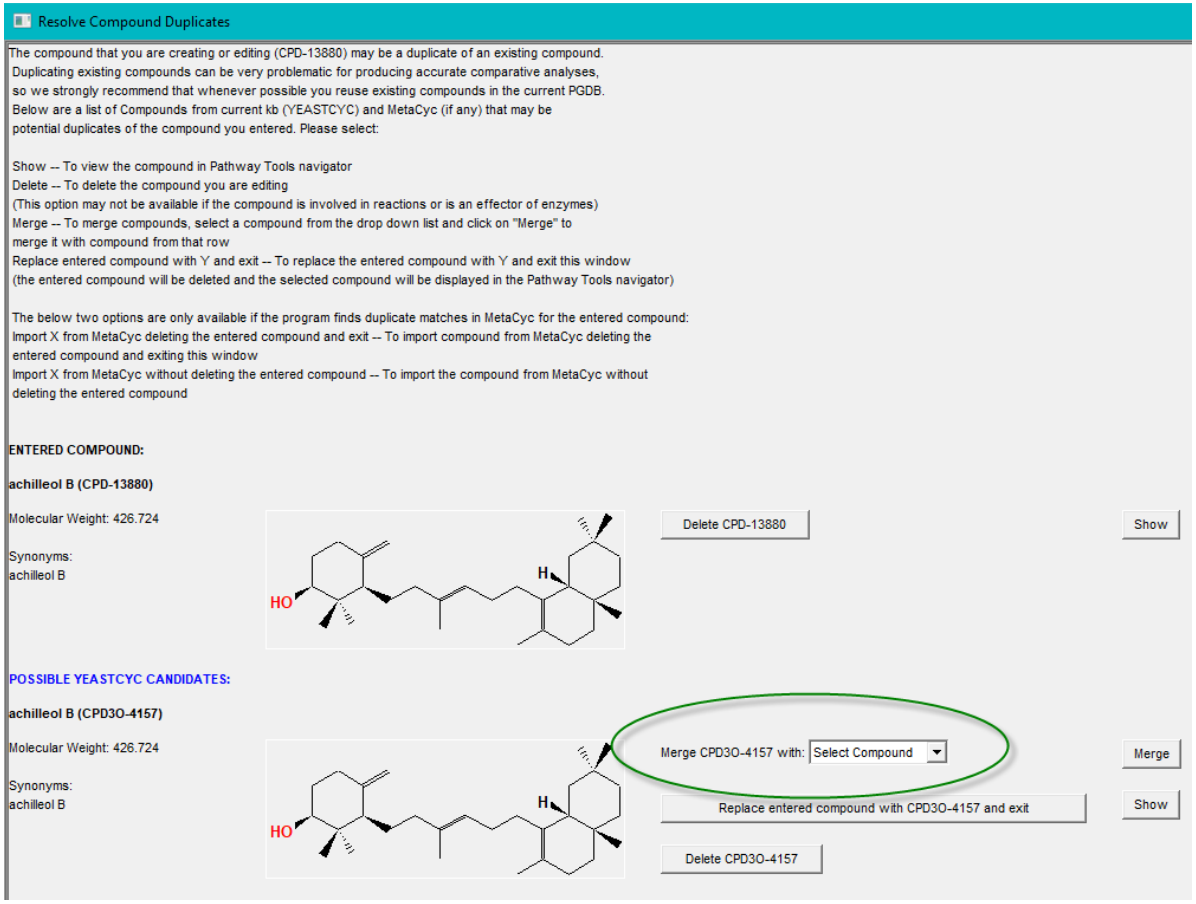
Compound duplicate checker - import

If the compound duplicate checker catches a duplicated compound upon its creation, you can delete the new compound and import the MetaCyc compound instead.



Compound duplicate checker - merge

If the compound duplicate checker catches an existing duplication, you can merge the two compounds, keeping the MetaCyc Frame ID.

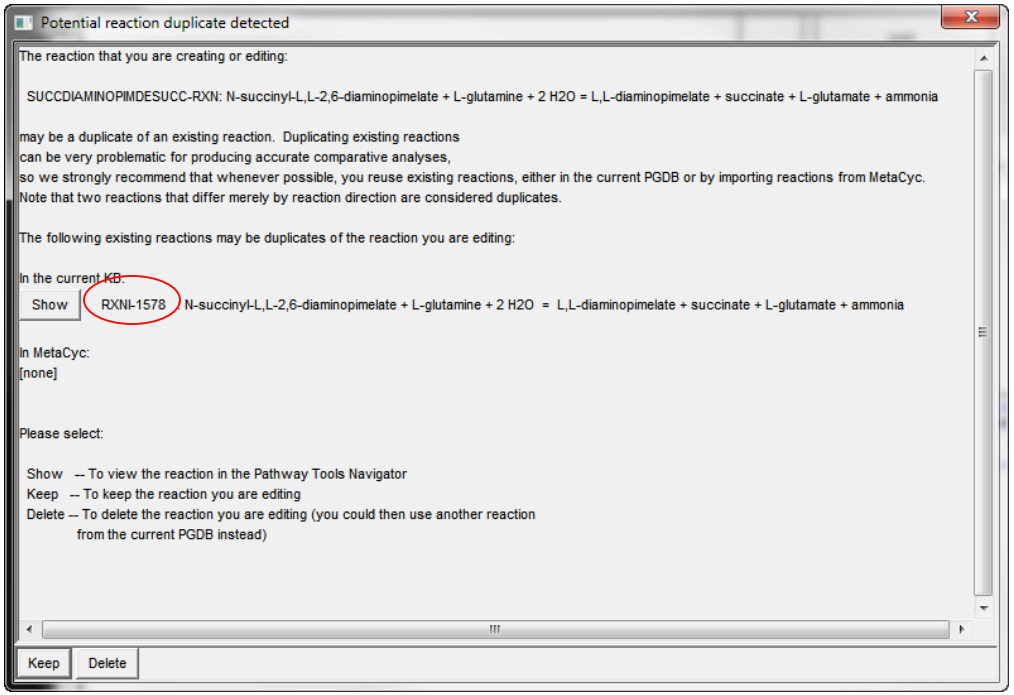


Reuse existing reactions instead of creating duplicates!

If the reaction is already present in your PGDB, you will see a window like this one

You should choose the option "Delete".

If you want to use the reaction in a pathway, press "show", then copy the frame ID of the existing reaction before you press Delete and close this window, so you could use it later when specifying the pathway.

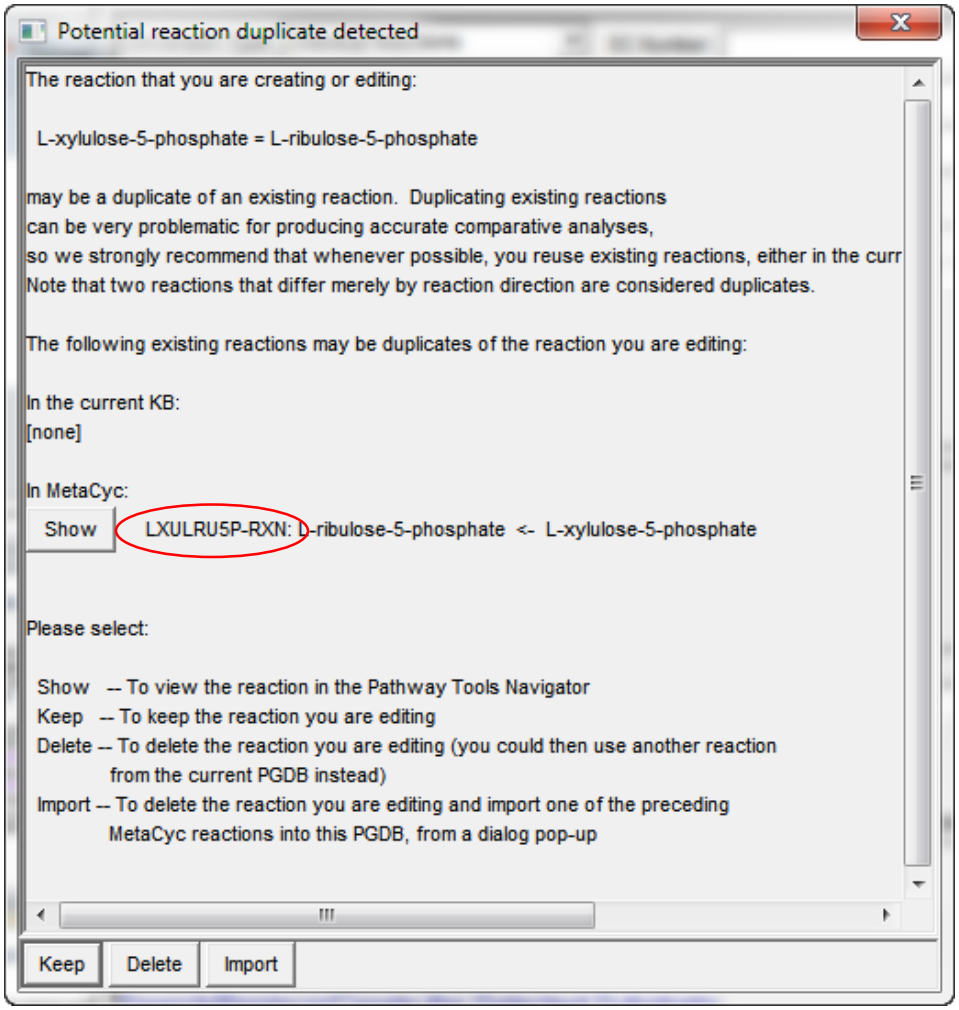


Import MetaCyc reactions instead of creating duplicates!

If the reaction is already present in MetaCyc (but not the current PGDB), you will see a window like this one

You should choose the option "Import".

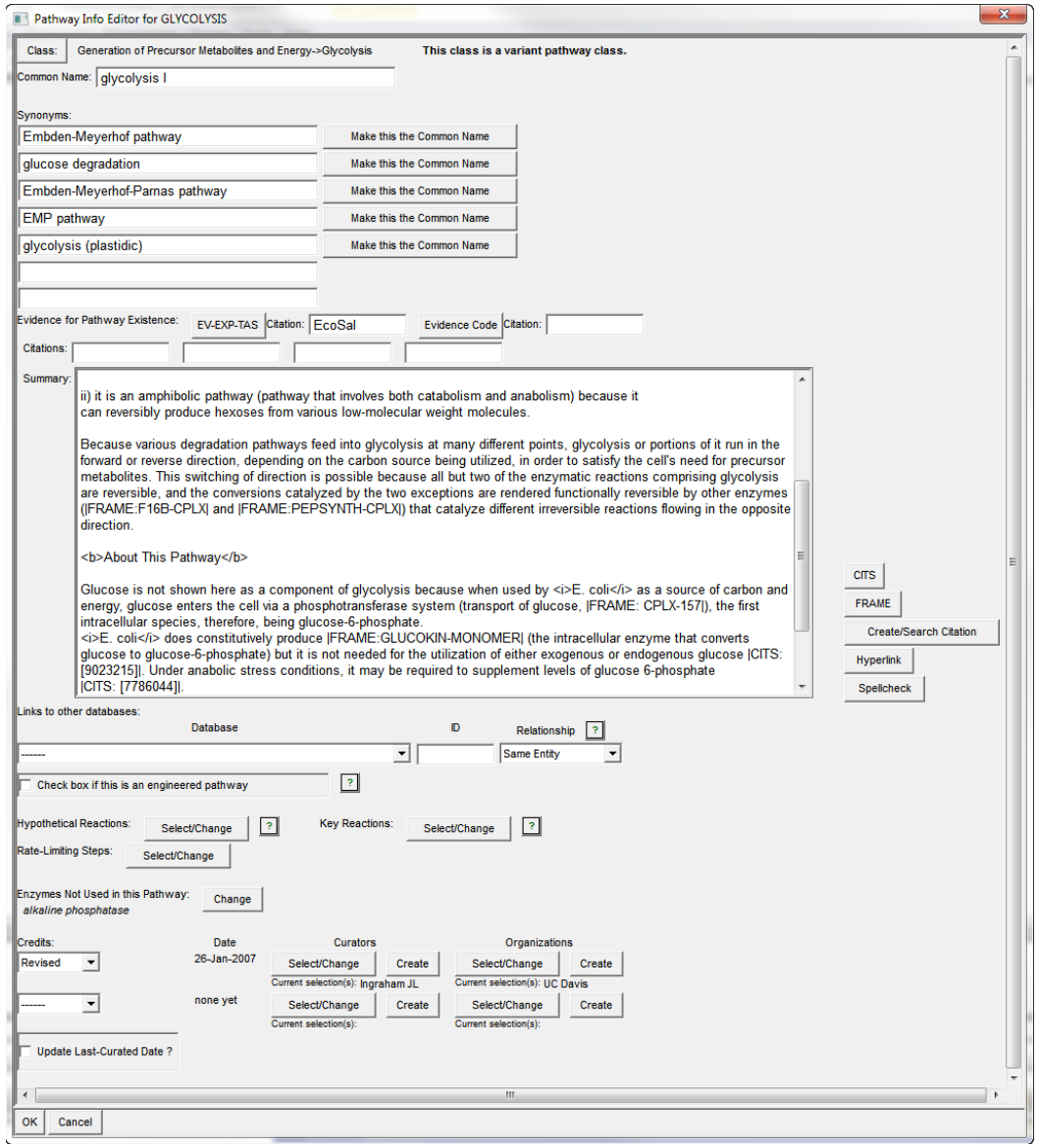
If you want to use the reaction in a pathway, copy the frame ID of the existing reaction before you press the Import button, so you could use it later when specifying the pathway.



PATHWAYS

The pathway info editor

- Class (variant class)
- Common Name
- Synonyms
- Evidence Codes
- Citations
- Summary
- External Links
- Hypothetical reactions
- Key reaction
- Rate-limiting steps
- Enzymes not in use
- Author credit

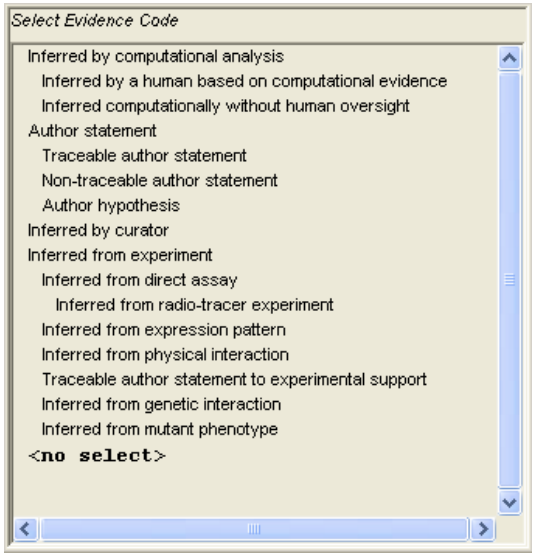


Evidence codes for pathways

http://bioinformatics.ai.sri.com/evidence-ontology/

Experimental evidence codes:

IDA: inferred from direct assay
IEP: inferred from expression pattern
IPI: inferred from physical interaction
TAS: traceable author statement
IGI: inferred from genetic interaction
IMP: inferred from mutant phenotype



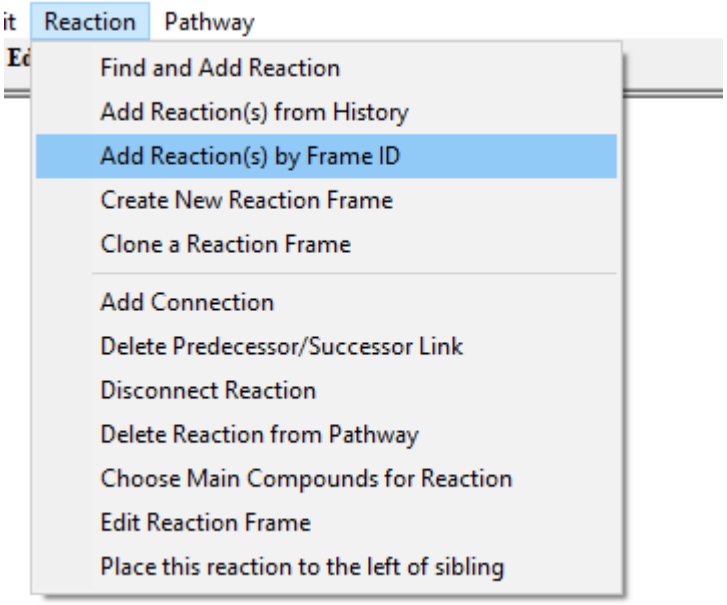
Full documentation for an evidence code is displayed in the Navigator (click the code icon)

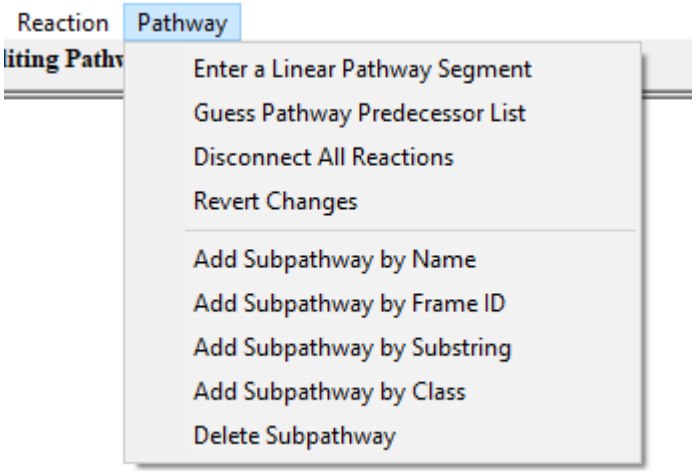
The pathway editor

Graphically create and modify pathways

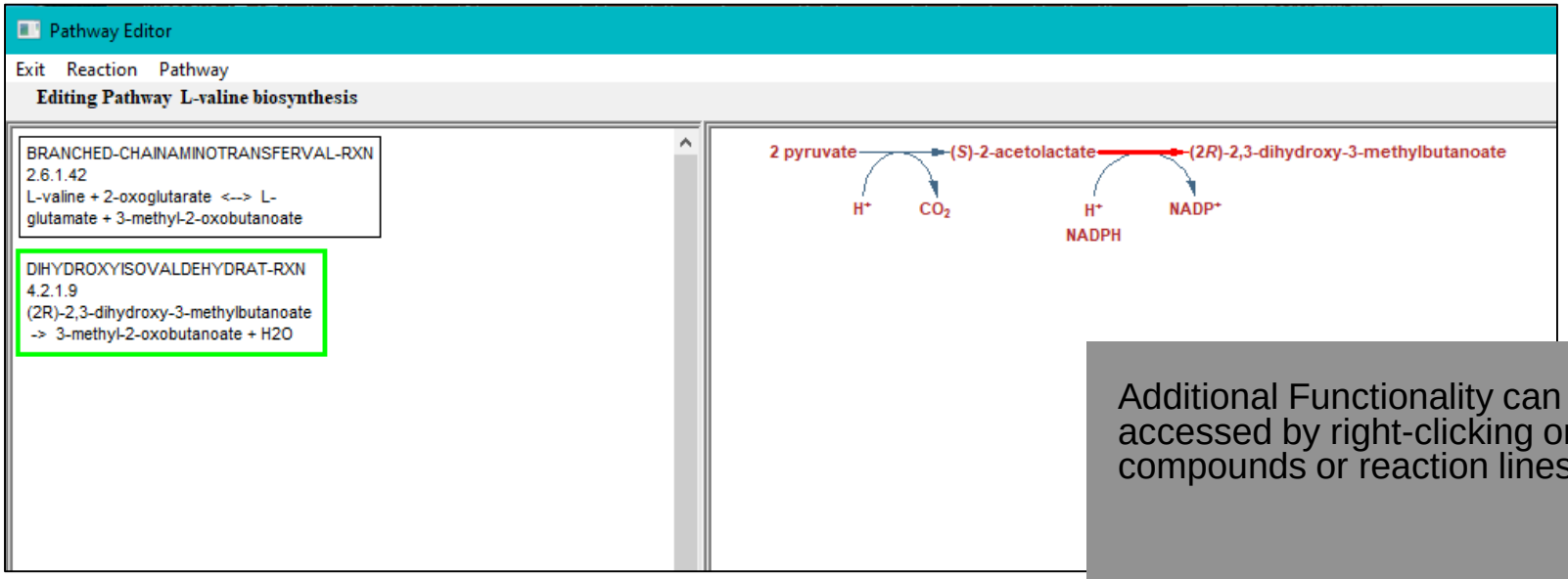
Reaction Menu: add reactions one by one

Pathway Menu: add sub-pathways to create a superpathway

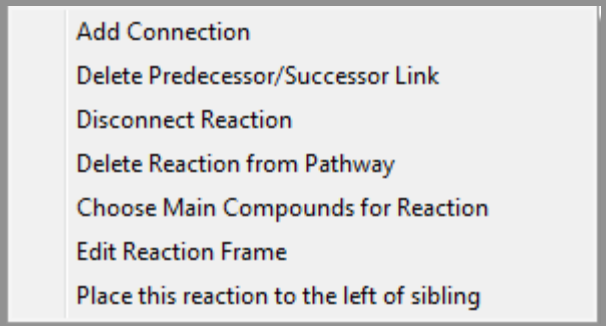




Connecting reactions

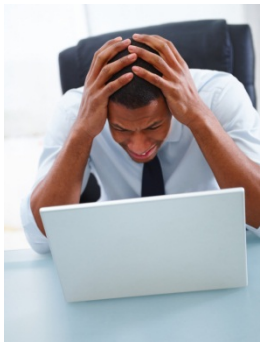


Additional Functionality can be accessed by right-clicking on compounds or reaction lines



(example: L-valine biosynthesis)

Pathway editor limitations

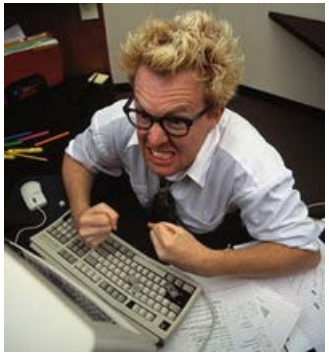


Complex situations can cause ambiguity:

- reaction directionality not specified
- reaction directionality opposite to direction in pathway
- dialog box for disambiguating
- pathway drawn in bizarre arrangement

- Fix:
  - try disconnecting reactions, specifying main compounds, and adding them in different order

Limitation: a reaction can appear only once in a pathway.



Homework

Please download “1. Reactions and pathways handout.pdf” from http://bioinformatics.ai.sri.com/ptools/tutorial/sessions/curation and complete the exercise before our next Zoom session.

If you run into difficulties or have any questions, contact me at ron.caspi@sri.com. I will be happy to answer any questions.