Editing Pathway/Genome Databases II

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

A lot more information is available in the Curator's Guide, at
https://bioinformatics.ai.sri.com/ptools/curatorsguide.pdf

Summary of last tutorial

What we discussed:

- Starting the software from the Lisp window
- Switching between the main window and Lisp window, breaks
- Data structure: frames, classes and instances
- Some menu commands
- Author credit system, creating curator and organization frames
- The compound, reaction, protein, and pathway editors
- Avoiding duplication

Summary of this tutorial

What we will discuss today:

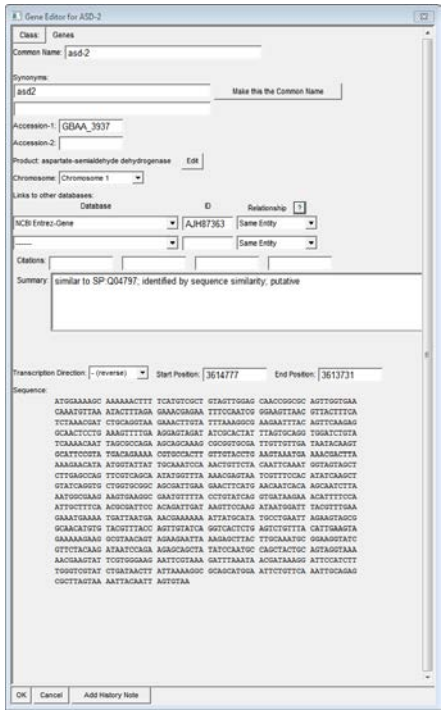
- Gene editor
- Creating protein complexes
- Writing summaries, using citations, internal hyperlinks
- Editing transcription units
- Curating regulatory interactions
- PGDB refinement operations
- Propagating MetaCyc updates
- The Consistency Checker

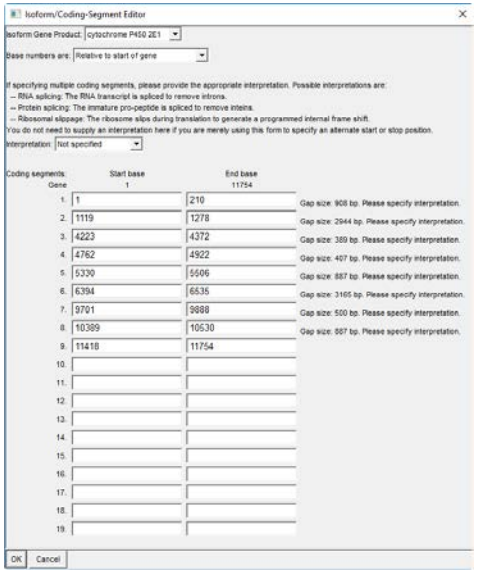
Once again, we will be using the PGDB for *Arthrospira platensis NIES-39* for practicing

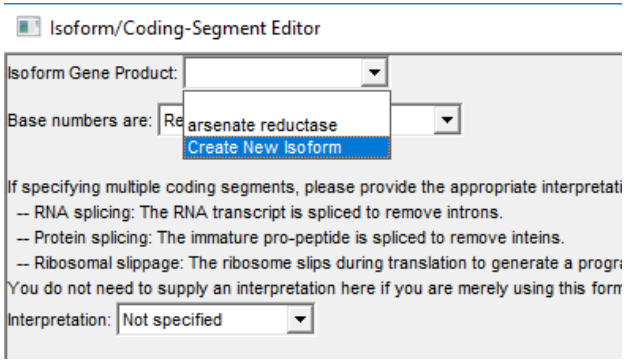
GENES

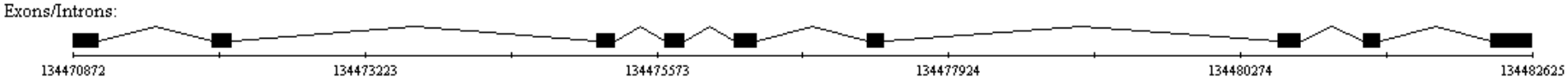
The Gene and Isoform/Coding-Segment Editors

- Enter synonyms/accession numbers
- Enter Links to other databases
- Define transcription direction
- Modify start/end positions
- Define introns
- Create new isoforms
- Define frame shifts









Adding a gene and its encoded protein

Sometimes there is a need to add a gene that is not present in the annotated genome. For the sake of demonstration, we will delete NIES39_RS03695 and its product, and then recreate them.

- Gene → New gene
- Change classification from “unclassified” to “ORF” (if appropriate)
- Enter gene name (pptX)
- Enter transcription Direction and coordinates (forward; start 779,299; end 779,499
- Enter link to NCBI RefSeq: NIES39_A08000
-
- Protein → New
- Macromolecule type → polypeptide
- Gene → pptX
- Name → putative serine/threonine phosphatase
- Enter link to RefSeq: BAI88638.1

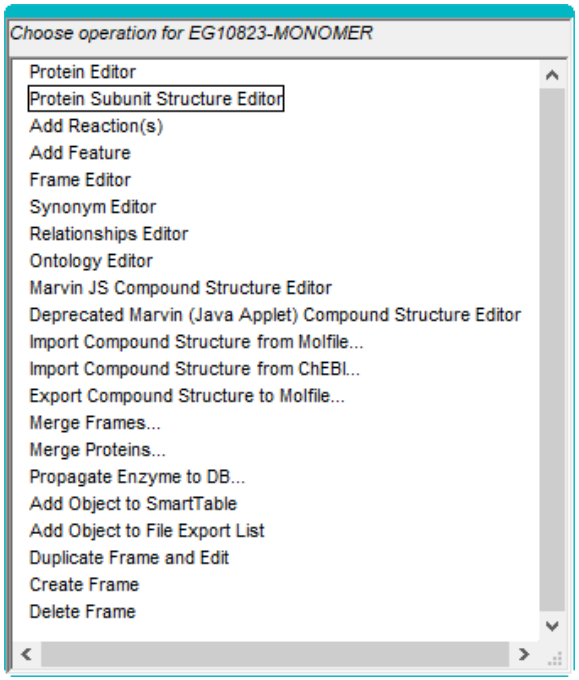
PROTEIN COMPLEXES

Creating Protein Complexes

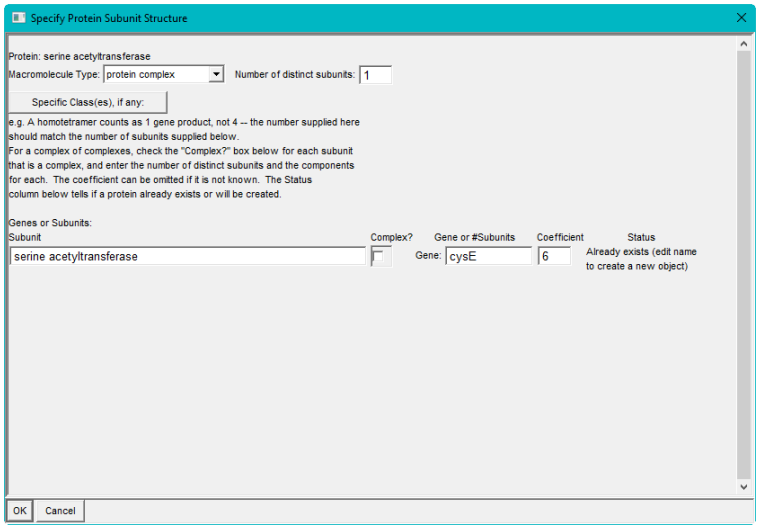
- Right-click on a protein and select

  Edit → Protein Subunit Structure Editor

- Change “Macromolecule Type” from polypeptide to protein complex



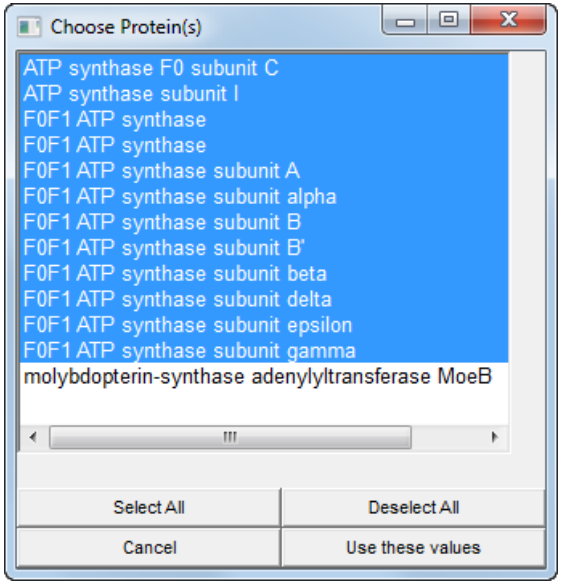
Example: a simple homohexamer

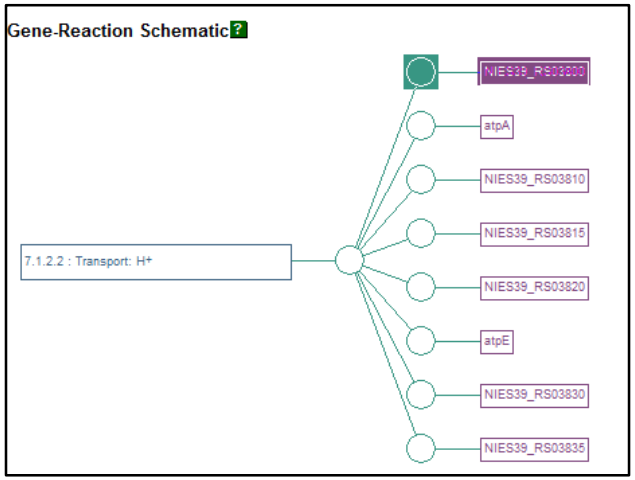


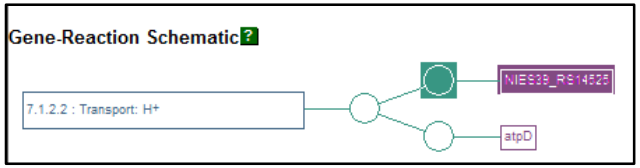
Example: defining ATP synthase

Search for ATP synthase subunits: Protein → Search by Substring → ATP synthase

We find 10 subunits arranged into two redundant complexes. Most genes do not have names.

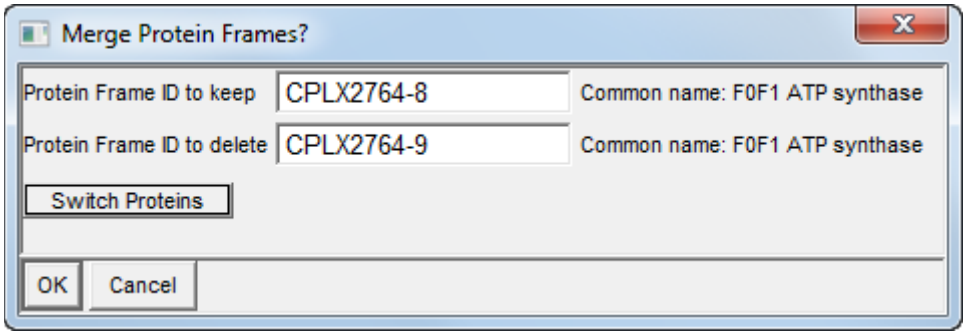






Merge the redundant complexes

- Copy the frame ID of the small complex:
  Right-click small complex → Show → Show frame ID
- Merge into the larger complex:
  Right-click large complex → Edit → Merge proteins



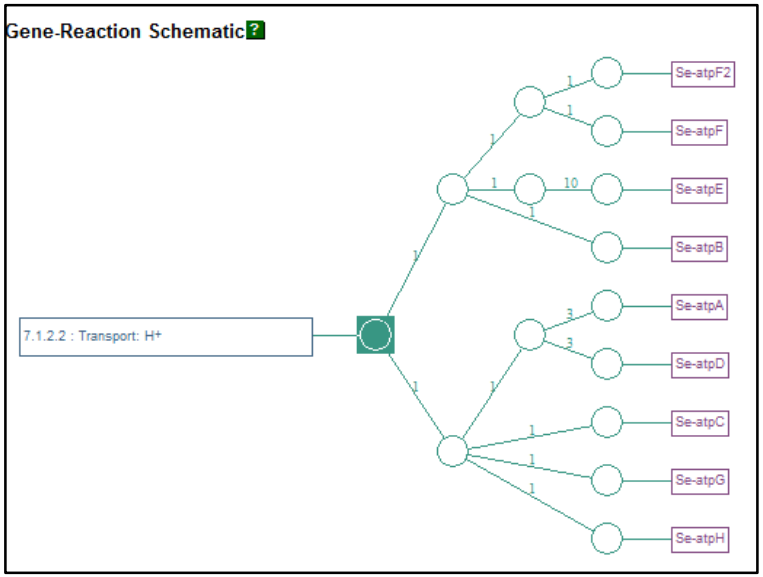
Get some help from MetaCyc

Right click on the reaction and select “Show frame in MetaCyc”

Click on the *Synechococcus elongatus* enzyme

Find out the proper gene name for the different subunits and enter them in the *A. platensis* PGDB

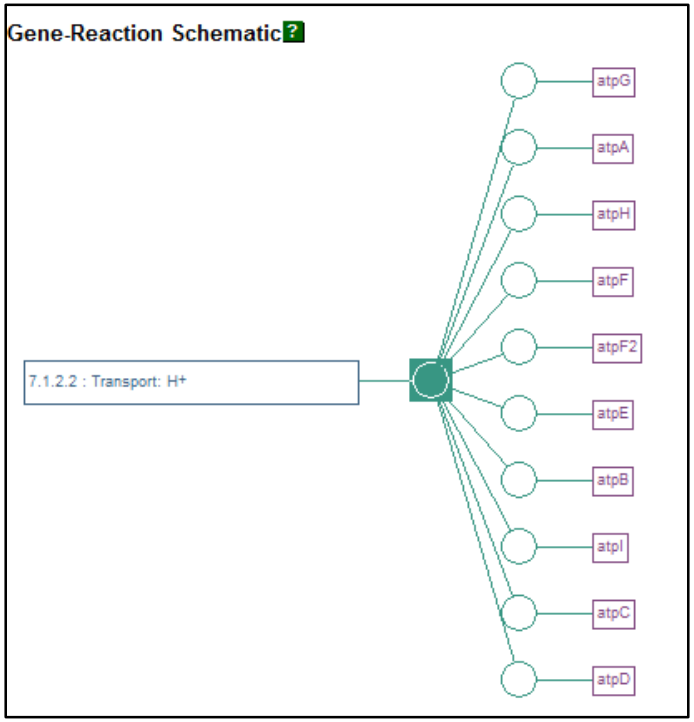
Subunit	**gene name**
(F_o sub-complex)	
a	atpB
b	atpF
b’	atpF2
c	atpG
(F_1 sub-complex)	
α	atpA
β	atpD
γ	atpG
δ	atpH
ε	atpC



Correct AtpI annotation

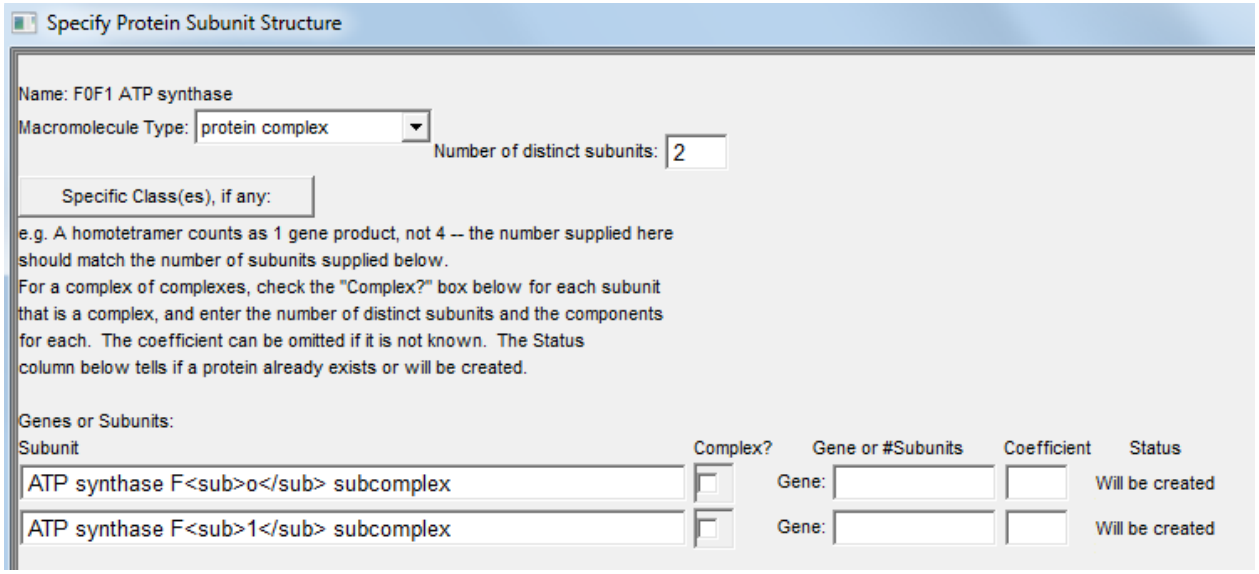
NIES39_RS03835-MONOMER was annotated as “ATP synthase subunit I”. However, you would notice that AtpI is not part of the MetaCyc complexes. AtpI has been characterized as an accessory protein.

- Change the gene name to *atpI* and the protein name to “ATP synthase accessory factor AtpI”



Defining the ATP synthase complex (I)

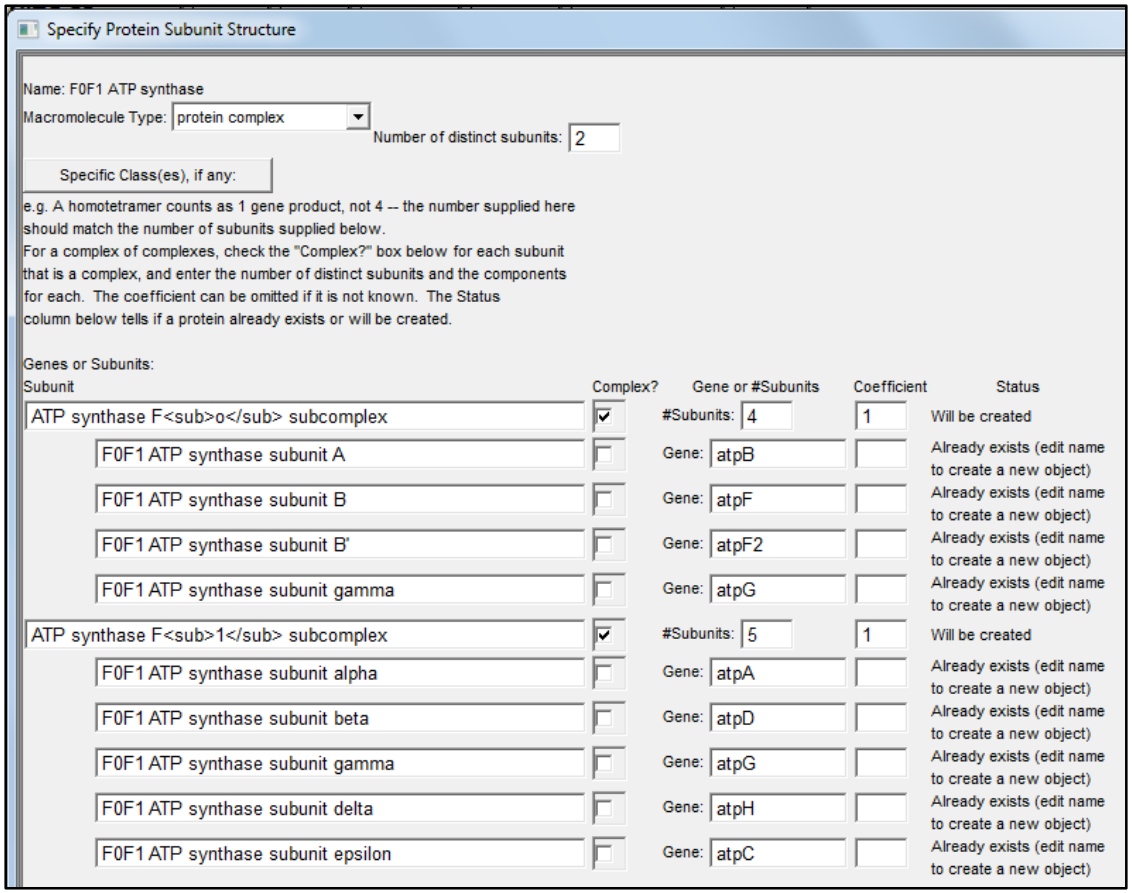
- Right-click on the complex → Edit → Protein Subunit Structure Editor
- Change number of distinct subunits to 2, and clear remaining fields.
- Enter names for the F_o and F_1 sub-complexes



Defining the ATP synthase complex (II)

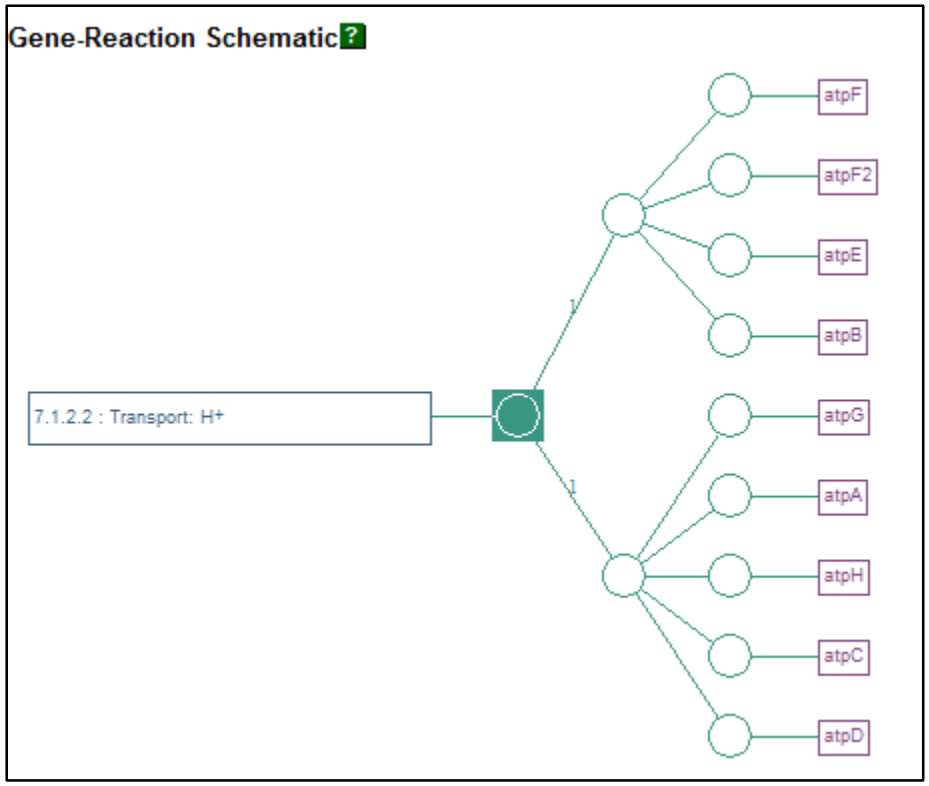
- Mark the two sub-complexes as complexes using the check box.
- Specify 4 subunits for F_o and 5 subunits for F_1
- Enter the gene names
- Click OK

(for simplicity we will not worry about coefficients)



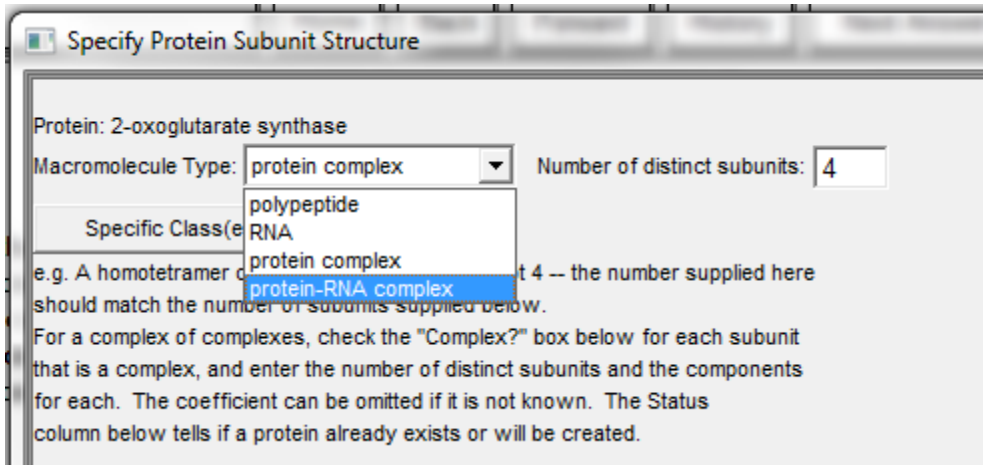
Complex is defined

Last thing to do is correct the name of the enzyme to F_oF_1 synthase and the activity name to simply ATP synthase



RNA/Protein-RNA complexes

To create a complex that contains either only RNA molecules or a combination of proteins and RNAs (e.g. a ribosome), simply select the appropriate type of molecule from the “Macromolecule Type” field of the “Protein Subunit Structure” editor



Using citations

- Most editors have citation boxes
- Open the ATP synthase complex in the protein editor
- Add a citation for PMID 33890627

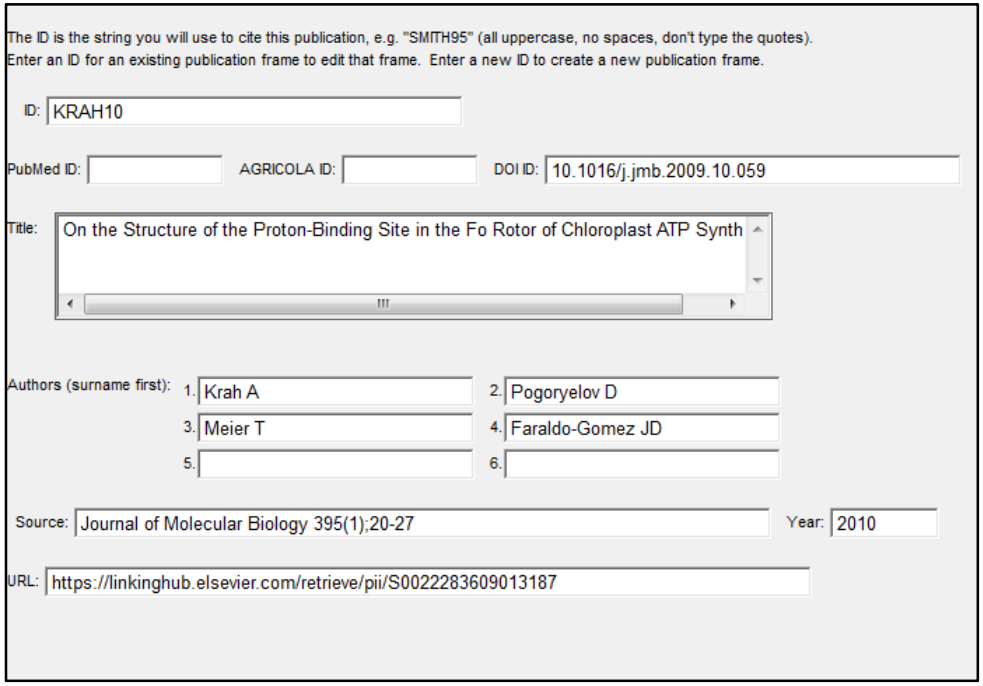
PubMed citations

- Paste the PubMed ID into a citation box
- Pathway Tools automatically imports the citation when exiting the editor
- Write a summary: “ATP synthase found in the thylakoid membranes of photosynthetic organisms has some unique features not present in other bacterial or mitochondrial systems”
- Use the CITS button to move the citation to the summary

Using non-Pubmed citations

You can type Pubmed IDs directly into the summary using CITS. Other types of citations must be first entered into a citation box as described earlier. Once they have been created, they can be used in the summaries

- Enter an ID in the form Krah10 in a citation box, invoke editor by clicking out of the box. Click on “Search or Create Publication Frame”.
- If you have a DOI number, enter it and click outside the DOI ID box, and it will be retrieved automatically (e.g. 10.1016/j.jmb.2009.10.059)
- If there is no DOI, type in the details.
- Click OK to close the editor



To make corrections invoke the Publication Editor by right-clicking on a citation and selecting Edit → Publication Editor

Writing Summaries

By writing summaries for enzymes and pathways you can turn your PGDB into a resource that integrates and summarizes the current knowledge about your organism.

Summaries should incorporate references to the literature, using the CITS button.

Using internal hyperlinks to other database objects is extremely useful. See for example

https://biocyc.org/SYNWH8102/NEW-IMAGE?type=ENZYME&object=CPLX1YI0-83

Using Internal Hyperlinks

- Internal hyperlinks are entered by using the |FRAME| button
- Clicking the button opens a list of the items in the History. Select the right one, and the hyperlink is created
- If the History list becomes too long, you can reset it by selecting Tools → History → Clear
- Alternatively, you can skip the history list by clicking the FRAME button a second time, which will insert an empty link. Print the frame ID of an object to the lisp window (right-click → Show → Show frame ID), copy it to the clipboard, and past it into the link.
- To modify what is shown in the summary, use this format |FRAME: frame-ID “type text here”|.
- You can use hyperlinks to MetaCyc objects.

Example for a summary with hyperlinks

EC 2.7.2.4, aspartate kinase, catalyzes the committed step in the pathways that ultimately lead to the synthesis of the amino acids lysine, threonine and isoleucine.

|FRAME: EC-2.7.2.4|, catalyzes the committed step in the pathways that ultimately lead to the synthesis of the amino acids |FRAME: LYS|, |FRAME: THR| and |FRAME: ILE|.

EC 2.7.2.4, aspartate kinase, catalyzes the committed step in the pathways that ultimately lead to the synthesis of the amino acids **L-lysine**, **L-threonine** and **L-isoleucine**.

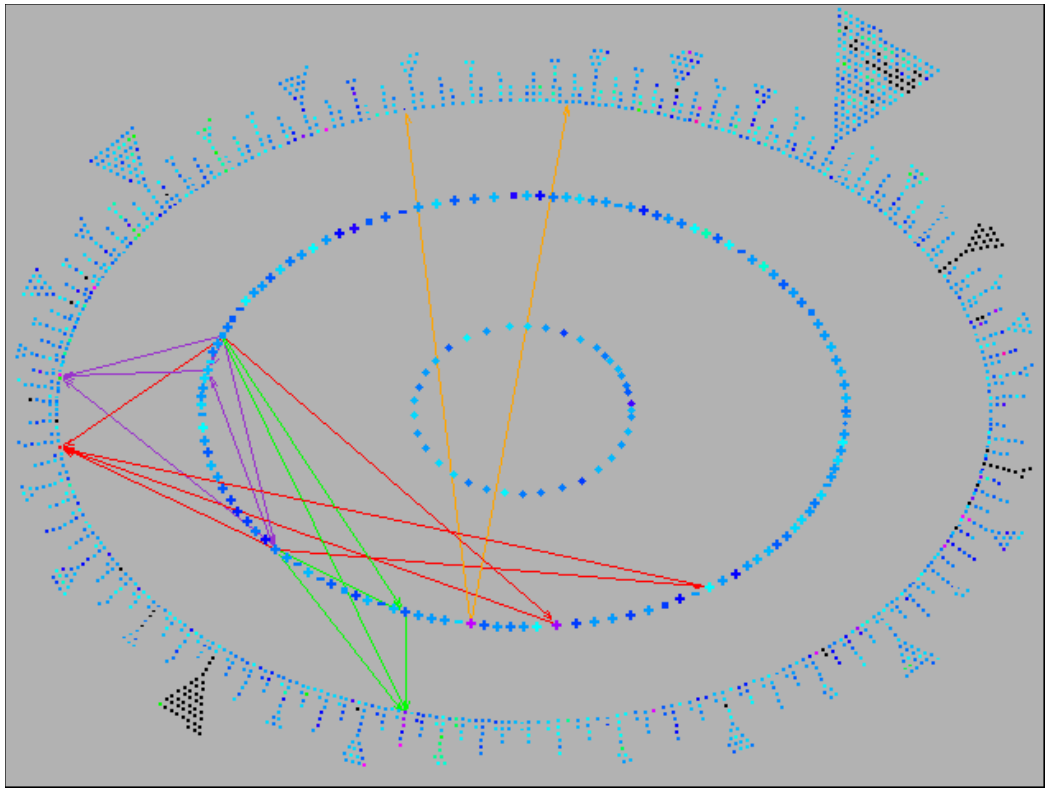
Copying MetaCyc comments

Pathways show up with the MetaCyc summary. If you want to modify a summary so it is applicable to your organism, you can copy the MetaCyc summary and alter it.

To copy a summary from MetaCyc:

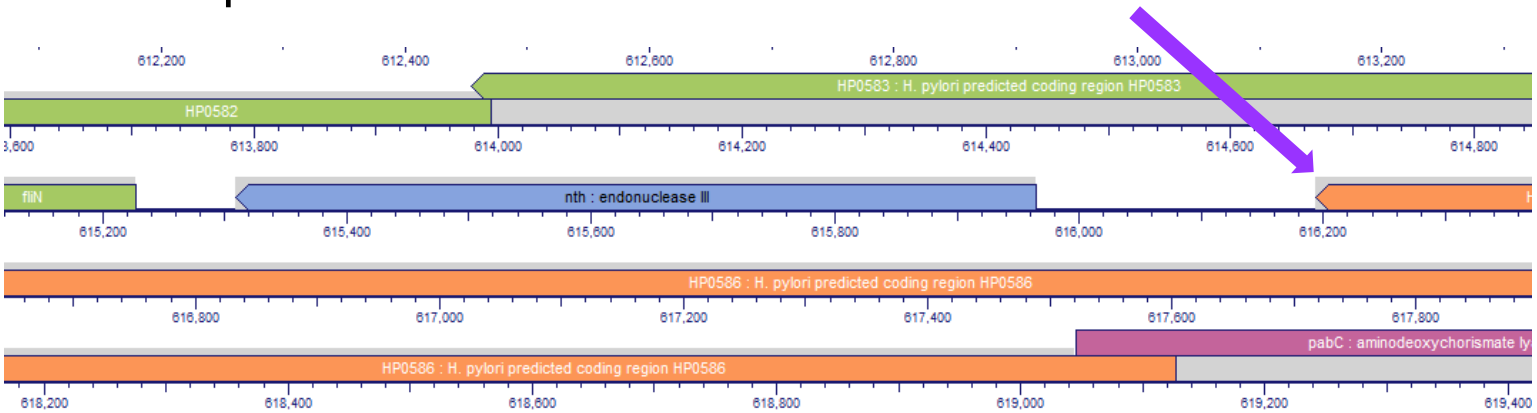
1. Type (dev) in the listener pane to activate MetaCyc editors
2. Right-click the pathway's name in your PGDB, select Show → Show frame in MetaCyc
3. Open the MetaCyc pathway in the Pathway Info Editor and copy the text
4. Go back to the pathway in your PGDB, open the editor and paste

Curating Regulation



Transcription Units

Transcription units are predicted by the PathoLogic tool “predict transcription units”.



Right-click the gray area and select Edit → Transcription Unit Editor. You can:

- Modify the included genes
- Add promoter information
- Specify sigma factor
- Add several different experimental evidence codes

Example for regulatory data

Let's make some (unbased) assumptions:

1. The *atpE, atpB*, and *atpI* genes are part of the transcription unit that contains the 5 genes upstream.

2. This TU is controlled by the sigma70 factor (NIES39_RS27630-MONOMER)

3. There is a promoter (atpIp) 85 bases before the atpI gene

4. Transcription factor (TF) NIES39_RS27155-MONOMER activates transcription, but only when bound to ADP

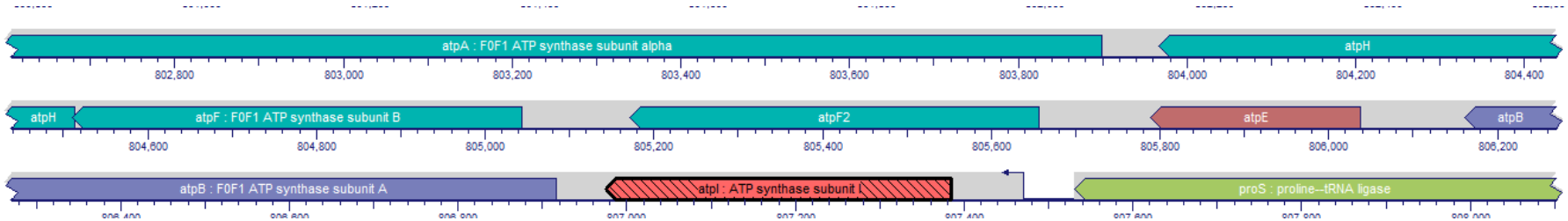
Modifying the transcription units

First we need to remove the incorrect TUs. Right-click on the Tus of *atpE, atpB*, and *atpI*, respectively, and select Edit → Delete frame (can be done from the gene page).

Next, right-click on the TU that needs expansion and select Edit → Transcription Unit Editor. Add the names of the three genes to the list of genes.

In the Promoter box type “atpIp”, and enter -85 in the box “distance from the first gene”.

From the list of sigma factors, select the sigma70 family factor.



Adding transcription factor (TF) interactions

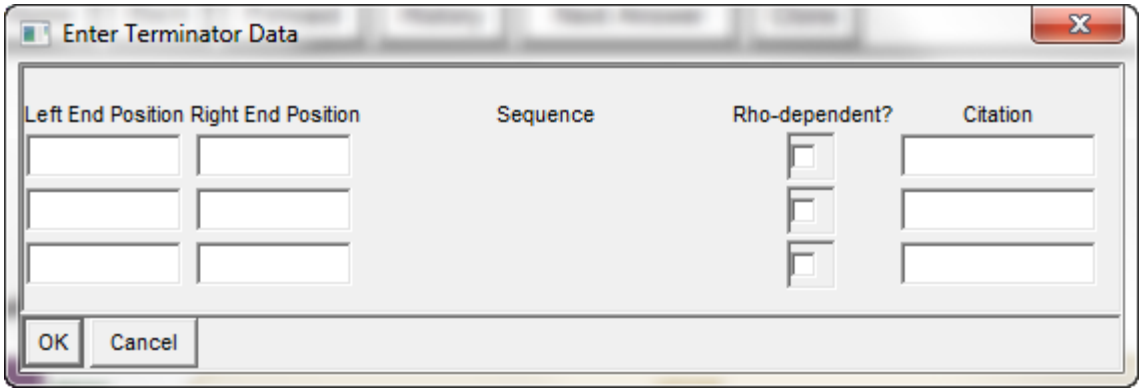
- Click on TU and select Edit → Create Regulatory Interaction
- In the window that pops up, select the atpIp promoter
- Paste the TF frame ID (NIES39_RS27155-MONOMER) in the “Protein” box
- Select “Activator” from the “Function” box
- Enter ADP for small molecule, and select “Active” in the box below
- Add evidence code and references
- Optional: define relative distance from transcription start site

- The information can be edited in the future by right-clicking on the green box and selecting Edit → Regulatory Interaction Editor

Specifying Terminators

Right-Click on the TU

Edit → Edit Terminators



More Regulatory Interactions

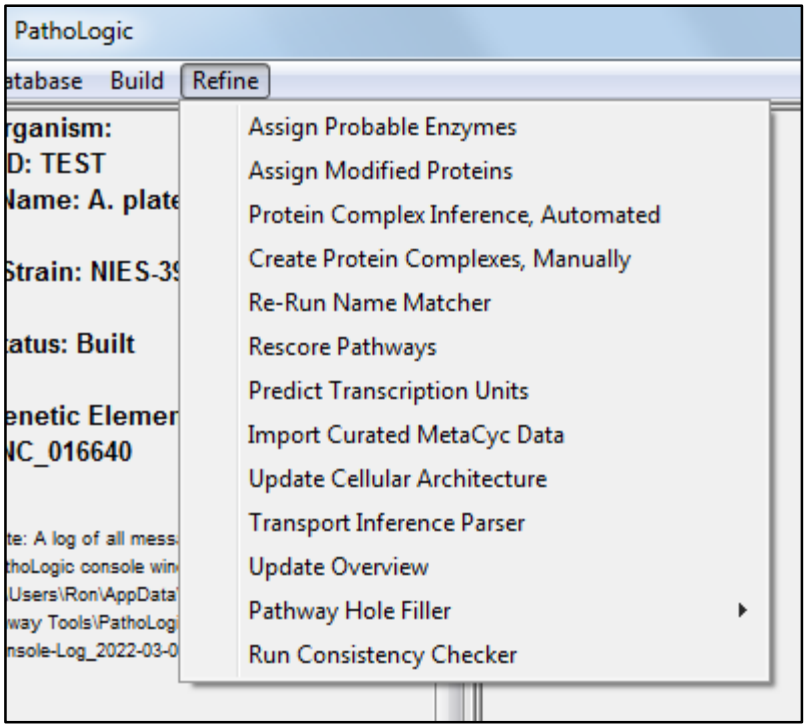
Pathway Tools supports additional types of regulatory interactions including:

- Attenuation
- Regulation of translation
  - RNA-mediated
  - Protein-mediated
  - Small molecule-mediated
- Regulated protein or mRNA degradation

When starting the Regulatory Interaction Editor, click the “Type of regulation” button at the top to select these types of interactions.

PathoLogic PGDB refinement tools

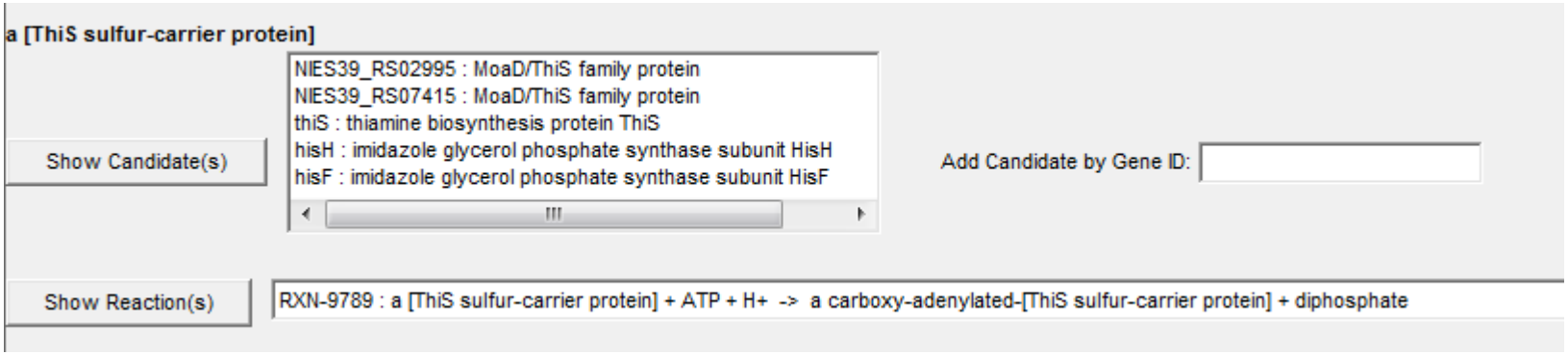
PathoLogic contains a number of tools that require curator interaction, that can add content and improve the PGDB quality



Assign modified proteins

In the previous webinar we discussed the fact that proteins that participate in MetaCyc reactions are all classes.

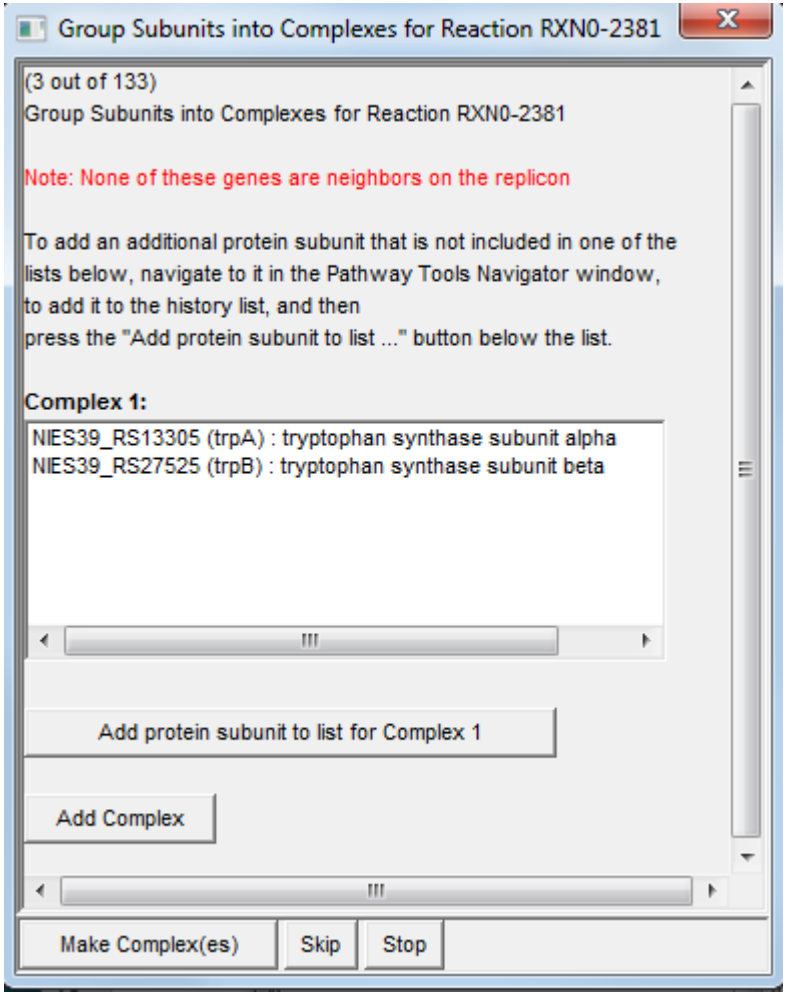
This tool enables linking the protein instances in the PGDB with these protein classes



Create protein complexes (manual tool)

Pathway Tools contains an automated protein complex creation tool. However, that tool only accepts complexes whose subunits are encoded by neighboring genes.

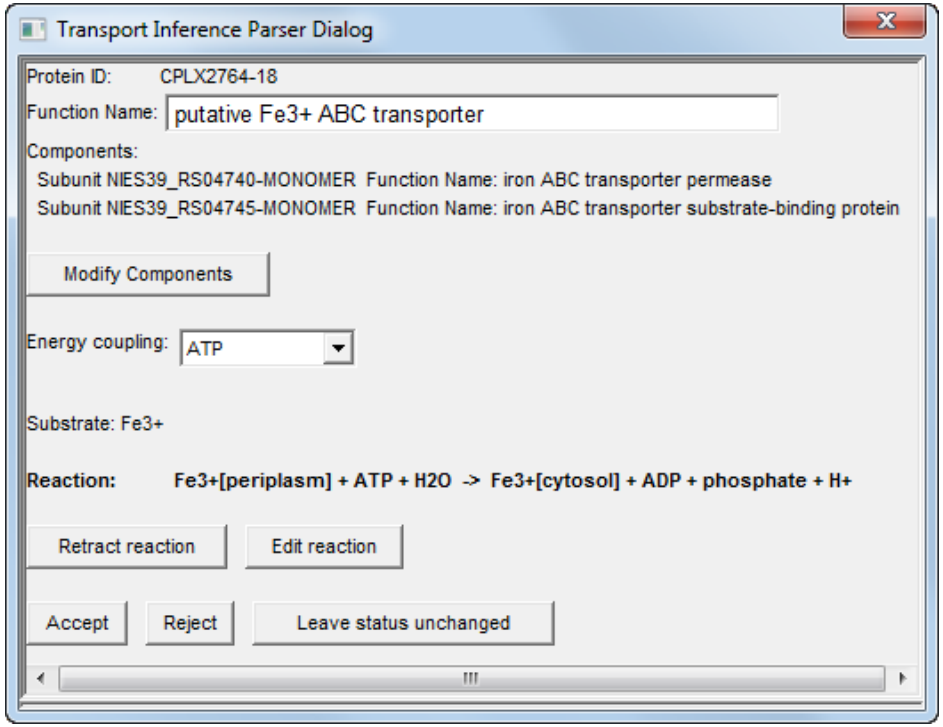
This manual tool allows a curator to inspect possible additional complexes and accept or reject them.



Transport Inference Parser (TIP)

TIP attempts to identify transporters in the genome based on their annotation and creates the most likely transport reactions that describes their activity.

It also creates complexes when it identifies a set of genes likely to encode a complex.



Pathway Hole filler (PHF)

The Pathway Hole filler attempts to identify missing enzymes in pathways based on sequence similarity to enzymes that have been assigned to these reactions in MetaCyc. This can overcome shortcomings in the annotation pipeline.

The PHF runs in three steps.

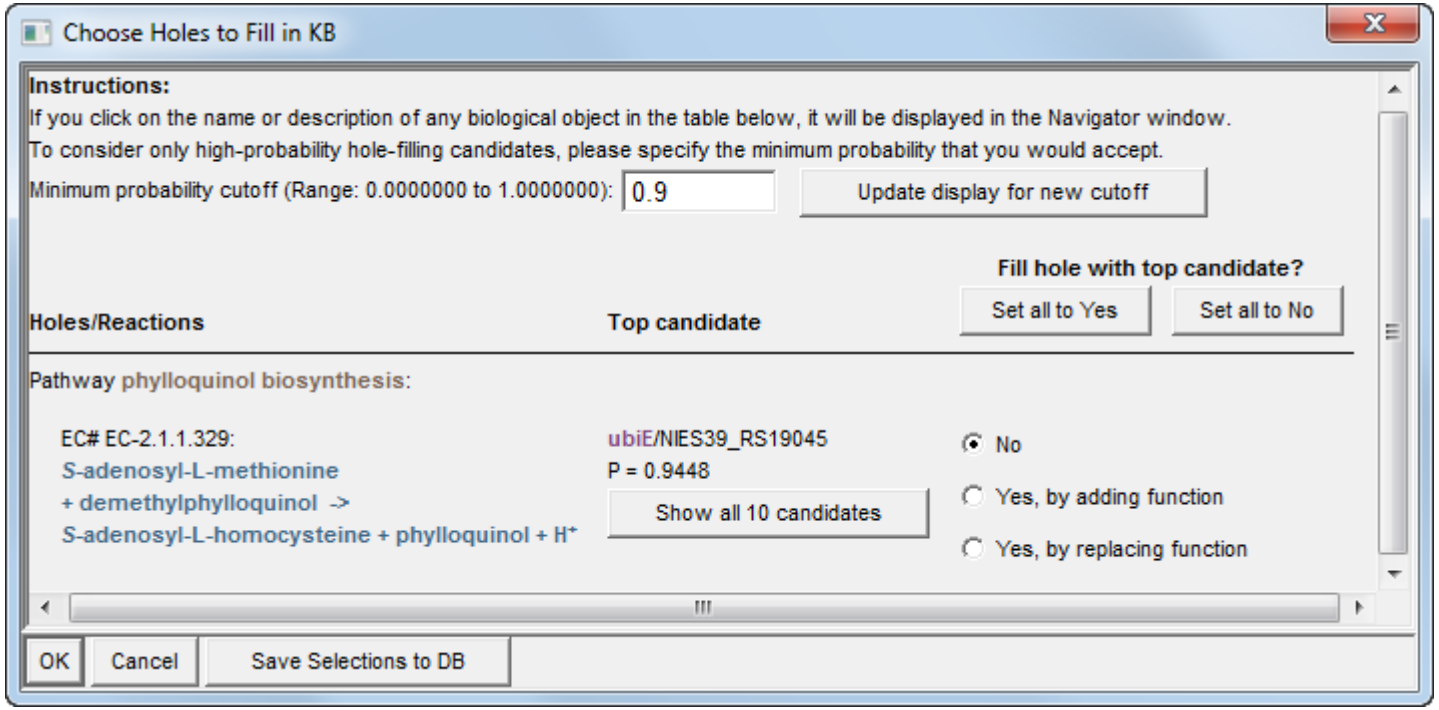
Step 1: Identify pathway holes, gather sequence data and build a training data set (takes 10-20 minutes) -automatic

Step 2: Identify candidate genes – automatic

Step 3: This is the step where the curator needs to go over the suggestions and decide which ones to accept - manual

PHF example

The tool correctly identifies a missing enzyme for the phylloquinol biosynthesis pathway, which has been misannotated.



And now, brought to you by Chitty Chitty Bang Bang and SRI International...

If your PGDB started out all smooth and shiny…

…but now it looks like this

...then it's time for an overhaul!

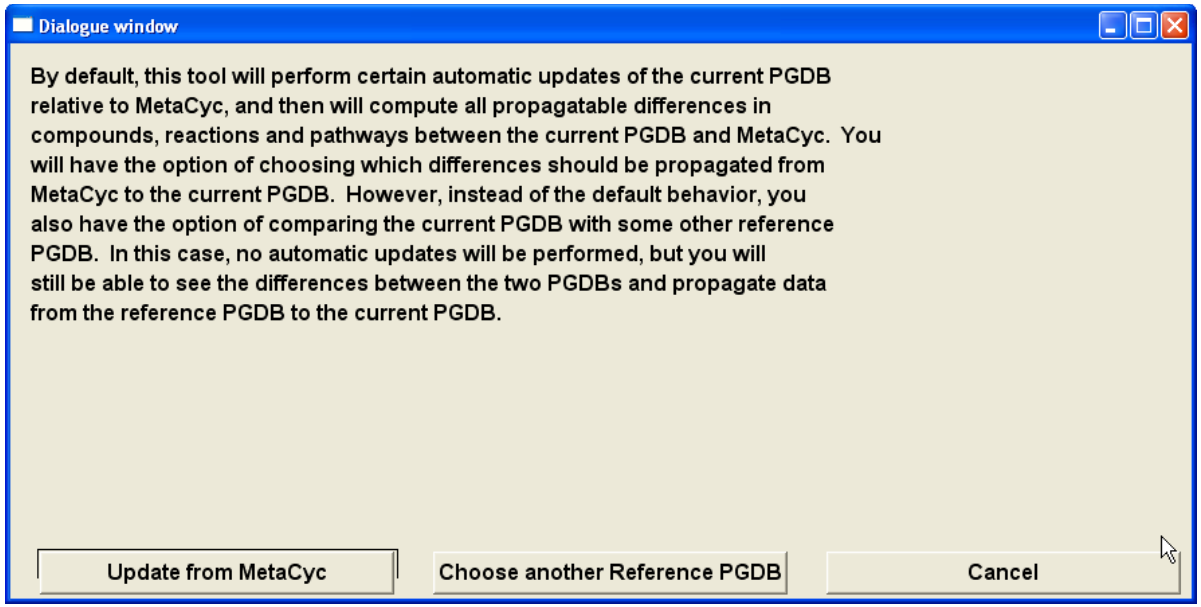
Tools for updating an aging PGDB

- Propagate updates from Reference DB (MetaCyc)
- Re-run the name matcher
- Rescore pathways
- Run the consistency checker

Propagating updates from a reference PGDB

Invoke from the Tools menu (Propagate MetaCyc Data Updates)

If your PGDB was created using a different reference PGDB, you can select it instead of MetaCyc



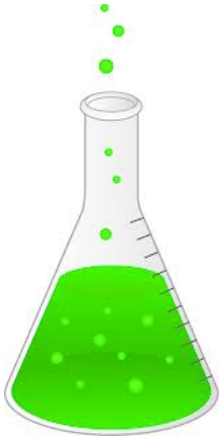
Propagating data updates

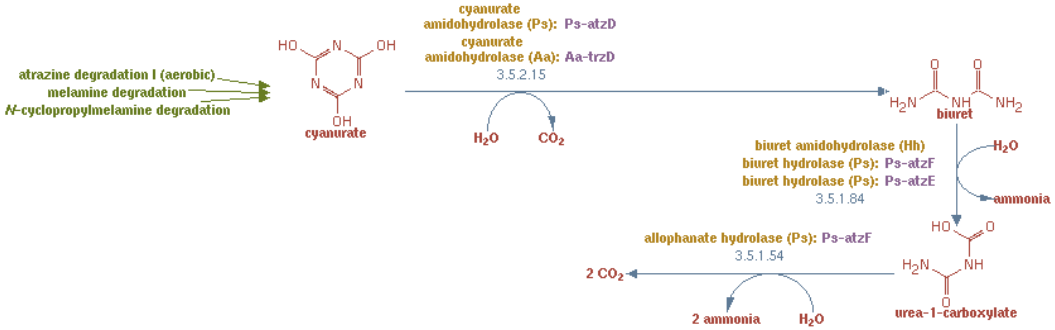
Data updates are broken into three sections:

Compounds

Reactions

Pathways





Propagating compound data

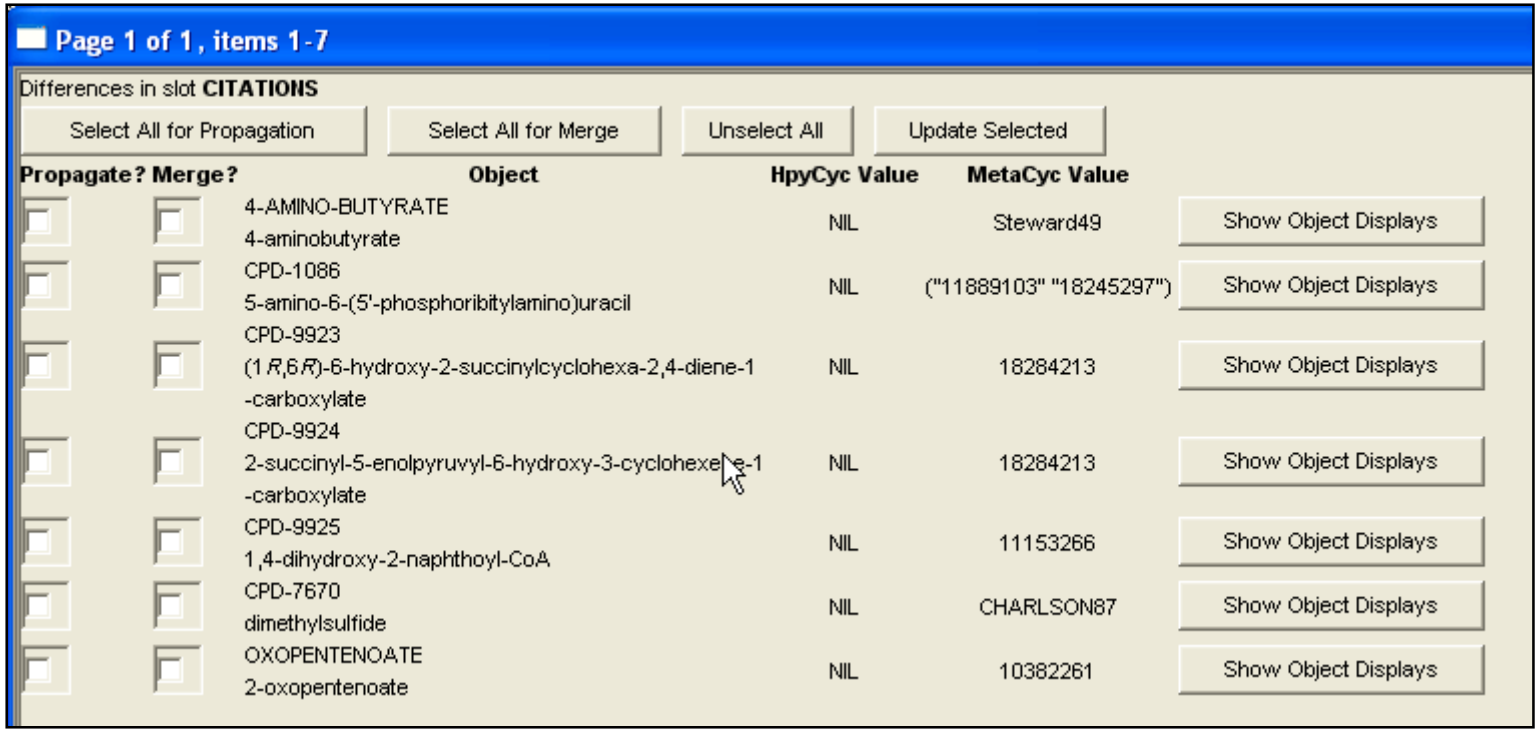
For compounds, the software looks for differences in chemical structures and in the data stored in the different slots

Compounds

33 Compounds have structures in MetaCyc but not in HpyCyc.	Select for Update	Propagate All	
537 Compounds have structure differences between MetaCyc and HpyCyc.	Select for Update	Propagate All	
5 Compounds have differences in slot N+1-NAME between MetaCyc and HpyCyc.	Select for Update	Propagate All	
5 Compounds have differences in slot N-1-NAME between MetaCyc and HpyCyc.	Select for Update	Propagate All	
4 Compounds have differences in slot N-NAME between MetaCyc and HpyCyc.	Select for Update	Propagate All	
2 Compounds have differences in slot OVERVIEW-NODE-SHAPE between MetaCyc and HpyCyc.	Select for Update	Propagate All	
42 Compounds have differences in slot COMMON-NAME between MetaCyc and HpyCyc.	Select for Update	Propagate All	
7 Compounds have differences in slot CITATIONS between MetaCyc and HpyCyc.	Select for Update	Propagate All	Merge All
2 Compounds have differences in slot GIBBS-0 between MetaCyc and HpyCyc.	Select for Update	Propagate All	
540 Compounds have differences in slot DBLINKS between MetaCyc and HpyCyc.	Select for Update	Propagate All	Merge All
141 Compounds have differences in slot SYNONYMS between MetaCyc and HpyCyc.	Select for Update	Propagate All	Merge All
9 Compounds are present in HpyCyc but not in MetaCyc.	Examine		

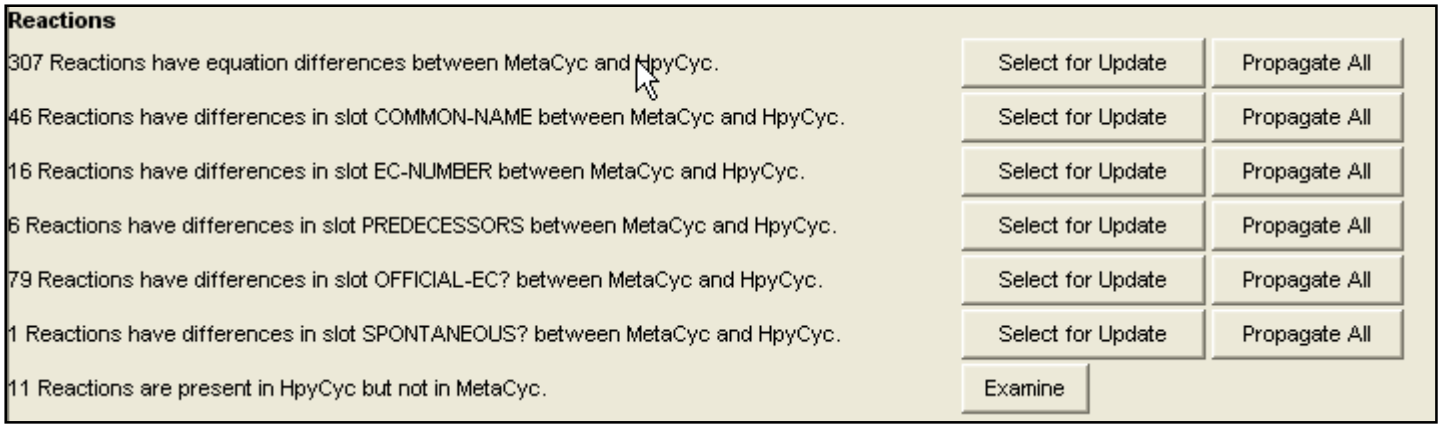
Inspecting differences

When you click the “select for update” button, you can review the differences and decide what to do for each case.



Propagating reaction data

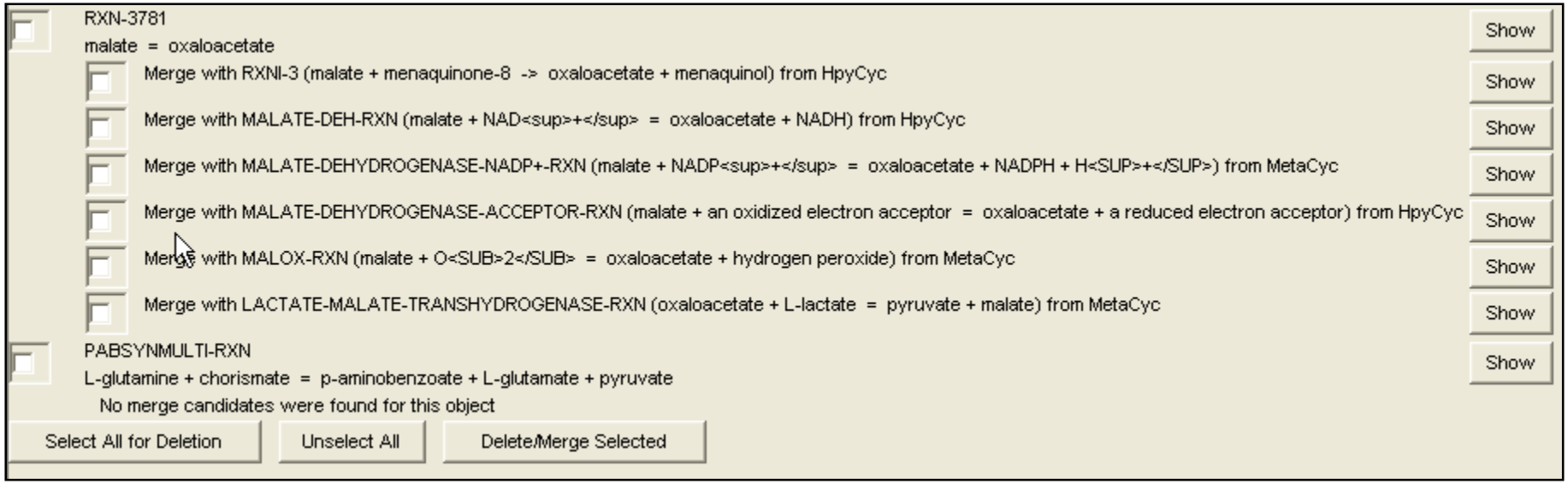
For reactions, the software looks for differences in the reaction equation, as well as in the data stored in the different slots



Objects not present in the reference database

When the software finds objects in the PGDB that are missing from the reference database, you can click the “Examine” button next to it to see the details.

The software would try to find merge candidates for these objects

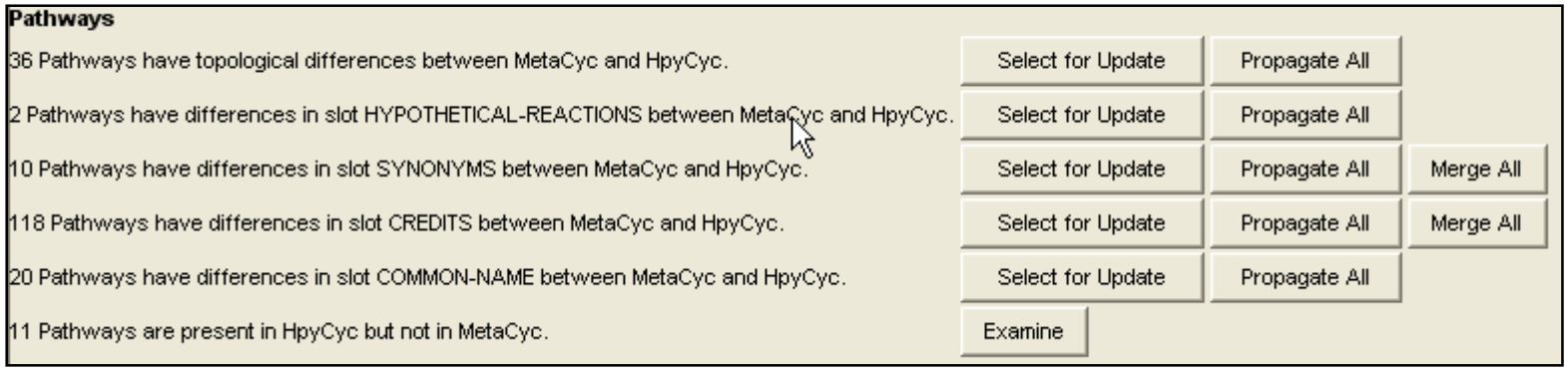


Propagating pathway data

For pathways, the software looks for differences in the topology of the pathway, as well as in the data stored in the different slots

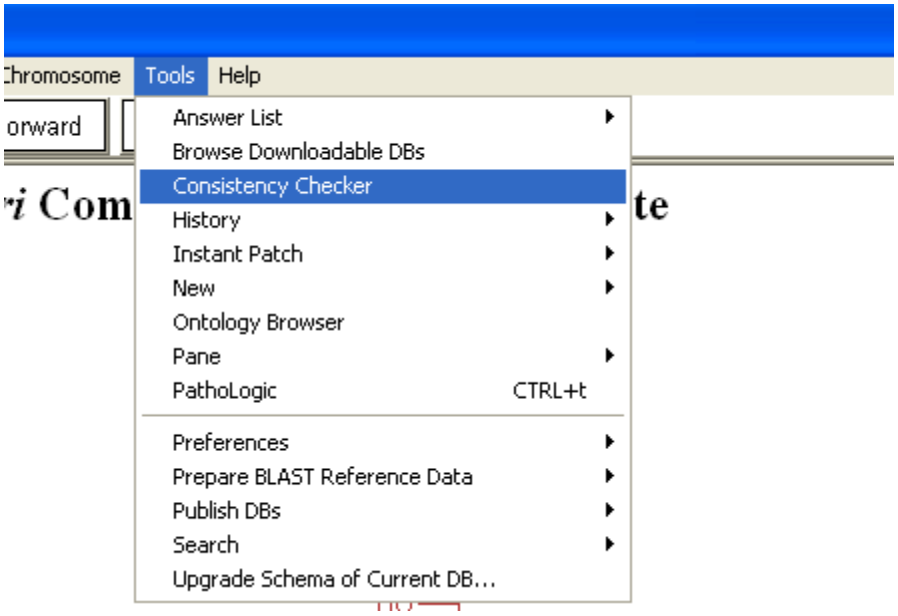
When pathways are present in your PGDB but not in the reference PGDB, it may be for two reasons: either you created them (in which case you would probably want to keep them), or they were deemed incorrect or redundant in MetaCyc, in which case you would want to delete them.

To make life easier: when modifying pathways in your PGDB, change the frame ID!

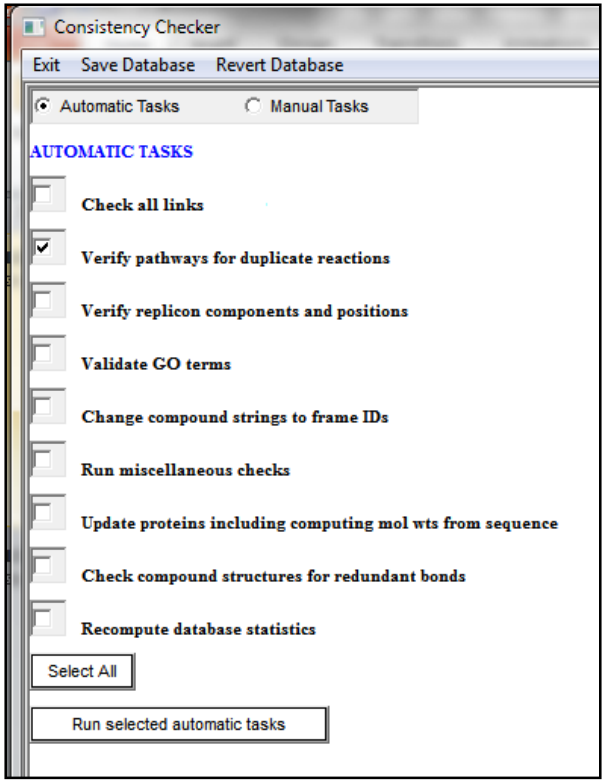


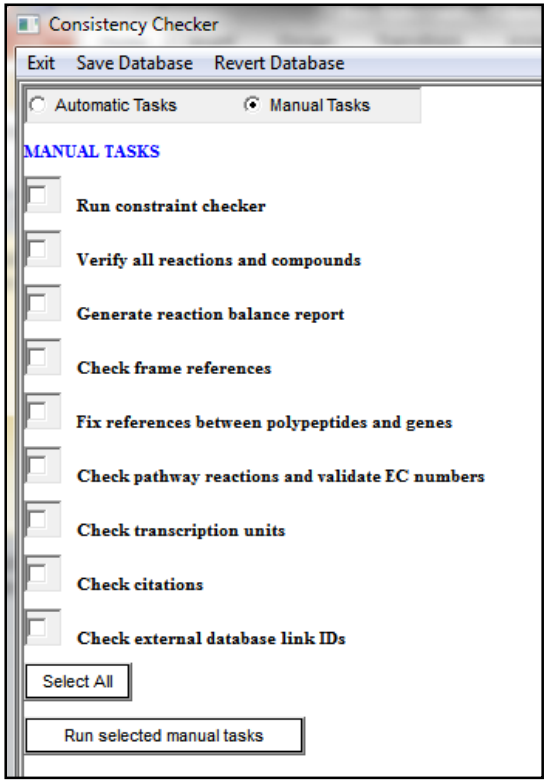
The consistency checker

Consistency checking should be performed routinely (every few months), and problems should be addressed



Automatic and manual tasks





- I recommend running the automatic tasks first
- I recommend running individual tasks one at a time.
- When you mouse over a task's name, you will see documentation for that particular task in the bottom window pane.

Consistency checker output

- The output appears on the right pane but is also saved into a text file in the reports directory. The name and location of the file are printed at the end of the output.

```
==Done checking all the links==

The report from this consistency checker run can be found at

C:\Program Files\Pathway Tools\ptools-local\pgdbs\registry\hpycyc\13.1\reports\consistency-checker-report-2009-08-13_11-24-56.txt
```

Automatic tasks: check all links

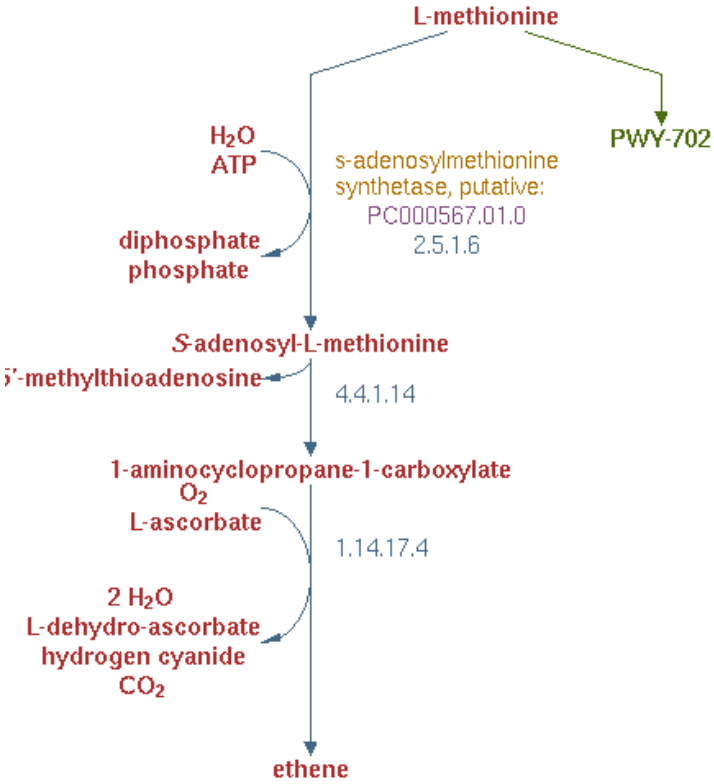
This tool looks at:

Inverse links (compound-reaction, gene-protein, etc.)

Pathway links

Ghost reactions in pathways

Pathways included in other pathways




```
===== Checking and removing any values from PATHWAY-LINKS that point to nonexistent frames ====

Removing link from pwy PWY-5901 to nonexistent pwys (ENTBACSYN-PWY)
```

Automatic tasks: check all links

Warnings are not necessarily errors but should be checked.

For example, PWY-21 is completely redundant to P142-PWY and should probably be deleted.

Warning:**MET-SAM-PWY** is completely contained within **PWYI-3** but is not listed in the SUB-PATHWAYS slot

Warning:**P142-PWY** is completely contained within **PWY-21** but is not listed in the SUB-PATHWAYS slot

Warning:**PWY-5600** is completely contained within **PWY-21** but is not listed in the SUB-PATHWAYS slot

Warning:**GLYCOLYSIS** is completely contained within **ANAEROFRUCAT-PWY** but is not listed in the SUB-PATHWAYS slot

Warning:**PWY-5485** is completely contained within **FERMENTATION-PWY** but is not listed in the SUB-PATHWAYS slot

Warning:**PWY-21** is completely contained within **P142-PWY** but is not listed in the SUB-PATHWAYS slot

Warning:**PWY-21** is completely contained within **PWY-5600** but is not listed in the SUB-PATHWAYS slot

Warning:**PWY-5484** is completely contained within **GLYCOLYSIS** but is not listed in the SUB-PATHWAYS slot

More automatic tasks

- Verify pathways for duplicate reactions
- Verify replicon components and positions: ensures all genes exist, sorts based on position.
- Validate GO terms: updates the GO terms using the latest version of GO-KB, removes obsolete ones.
- Change compound names to string IDs: mostly applies to legacy data, where enzyme regulators may have been entered as text strings.

Yet more automatic tasks

- Run miscellaneous checks: formatting glitches in names, validity of superpathways, clears values of computed slots, deletes temporary frames created by breaks when the pathway editor runs

- Update proteins: molecular weights recalculated from sequence

- Check compound structures for redundant bonds

Automatic tasks: recompute database statistics

Updates the numbers on the home page

Helicobacter pylori

Strain: 26695 HpyCyc version: 13.1

Generate Pathway Evidence Report | Generate Pathway Hole Report

Authors: Suzanne Pale, SRI International; Peter D. Karp, SRI International

Citations: [**Tomb, 1997**; **Marais, 1999**]

Replicon	Total Genes	Protein Genes	RNA Genes	Pseudogenes	Size (bp)
26695 Chromosome	1609	1566	43	0	1,667,867

Pathways:	143
Enzymatic Reactions:	671
Transport Reactions:	29
Polypeptides:	1598
Protein Complexes:	29
Enzymes:	330
Transporters:	33
Compounds:	553
Transcription Units:	817
tRNAs:	38

Manual tasks: the constraint checker

This tool usually requires the most time and effort for correcting the problems.

Flags constraints issues. For example, if a slot is supposed to contain only compound frame IDs, but a different type of frame is listed among its values, the constraint checker identifies and flags the offensive value.

The opposite is true as well: the checker will flag that compound as present in a slot of a frame that is not supposed to have such a value.

(this means errors are often listed multiple times, under different frames)

The checker also flags cardinality violations. For example, cases where more than one values are present in a slot that is only allowed to have a single value.

More manual tasks

- Verify all reactions and compounds: finds orphan enzymatic reaction frames (missing a protein, a reaction, or both); finds orphan reactions that are not associated with any other objects, looks for duplicate compounds.
- Generate reaction balance report

```
==== Reaction balance summary report for hpycyc ====

TOTAL BALANCED REACTIONS: 449
   With :CANNOT-BALANCE? slot set to TRUE: 0
TOTAL UNBALANCED REACTIONS: 46
   With :CANNOT-BALANCE? slot set to TRUE: 1
   With :CANNOT-BALANCE? slot not set: 45
TOTAL UNDETERMINED REACTIONS: 11
   With one or more of the substrates lack a chemical structure: 11
   With non-numerical coefficients: 0
```

Frame references errors

Frame AGMATHINE. is referenced in a |FRAME: | construct, but

does not exist either here or in MetaCyc or in EcoCyc. It is referenced in the

following places:

Frame: **PWY0-1299**
Slot: COMMENT

Looking at that pathway's comment, we find that the FRAME construct is missing the last bar.

ginine-dependent acid resistance system which couples
gmatine antiporter, AdiC, with arginine decarboxylase, AdiA.
nal |FRAME: ARG| for internal |FRAME: AGMATHINE. Arginine
ll arginine is decarboxylated by AdiA to agmatine, releasing
ith a proton. Agmatine is then exported through AdiC.

More manual tasks

- Fix references between polypeptide and genes: adds the gene value to modified proteins that miss it, adds a capitalized gene name to the synonyms list, scans that list for duplicates, flags orphan genes and proteins.
- Check pathway reactions and validate EC numbers: checks the PREDECESSORS slot of pathway frames, flags references to deleted and transferred EC numbers.
- Check transcription units: looks for invalid frames, transcription units with no genes, with genes in different directions, etc.

Even more manual tasks

- Check citations: tries to find formatting problems, downloads PubMed citations that have not been imported, provides statistics.
- Check external database link IDs: flags frames that are linked to the same external DB entry by links that are supposed to be unique.
- Check HTML tags: looks for formatting errors in HTML within comments.

And when you finish, take pride at your newly renovated PGDB!

Homework

Perform the following exercises at your pleasure

Exercise 2: Assigning enzymes and creating protein complexes

- Assigning enzymatic activities to proteins
- Defining protein complexes
- Creating a publication frame
- Exporting a pathway to a file