The Consistency Checker, or Overhauling a PGDB

By Ron Caspi

PGDB Atrophy

Your PGDB started out all smooth and shiny…

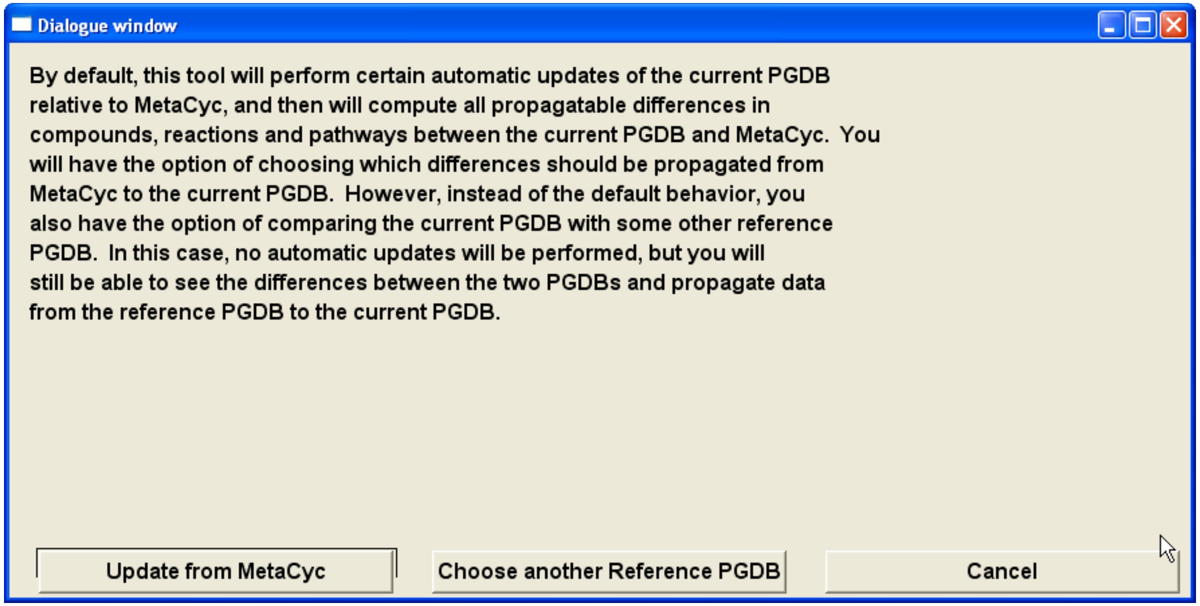
…but after a few years, it looks more like this

It’s time for an overhaul!

- Update genome annotation
- Propagate updates from Reference DB (MetaCyc)
- Re-run the name matcher
- Rescore pathways
- Re-run the transcription unit predictor
- Run the consistency checker
- Create protein complexes
- Re-run the Transport Inference Parser

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



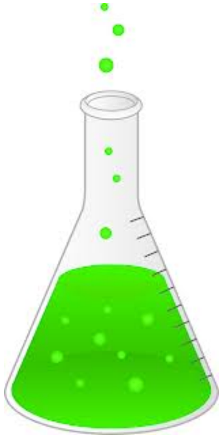
Dialogue window

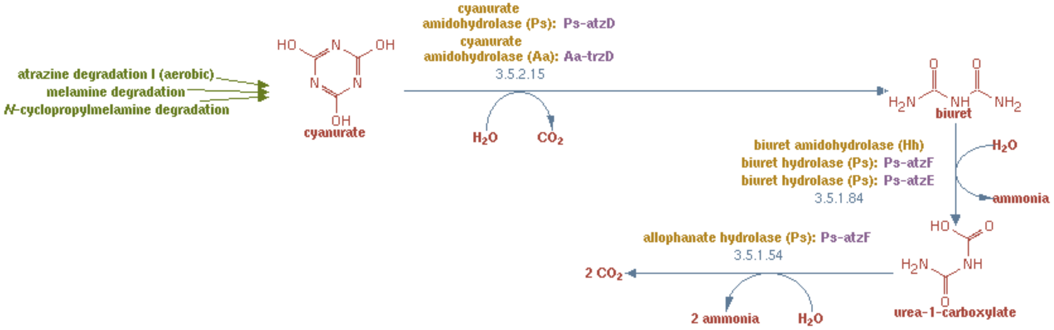
By default, this tool will perform certain automatic updates of the current PGDB relative to MetaCyc, and then will compute all propagatable differences in compounds, reactions and pathways between the current PGDB and MetaCyc. You will have the option of choosing which differences should be propagated from MetaCyc to the current PGDB. However, instead of the default behavior, you also have the option of comparing the current PGDB with some other reference PGDB. In this case, no automatic updates will be performed, but you will still be able to see the differences between the two PGDBs and propagate data from the reference PGDB to the current PGDB.

Propagating Data Updates

Data updates are broken into three sections:

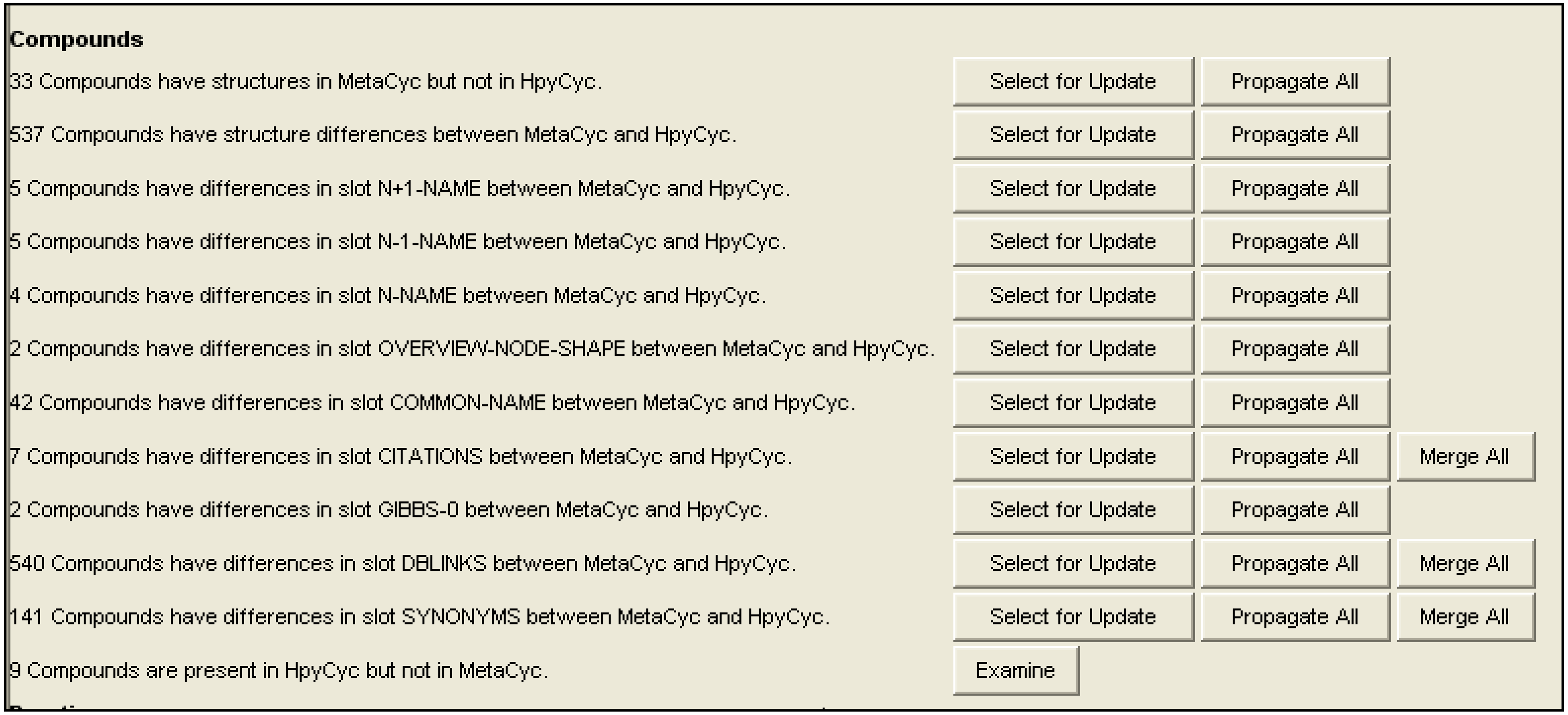
- Compounds
- Reactions
- Pathways




atrazine degradation I (aerobic)
melamine degradation
N-cyclopropylmelamine degradation
HO
N
OH
N
N
OH
cyanurate
cyanurate amidohydrolase (Ps): Ps-atzD
cyanurate amidohydrolase (Aa): Aa-trzD
3.5.2.15
H₂O
CO₂
O
O
H₂N
NH
NH₂
biuret
biuret amidohydrolase (Hh)
biuret hydrolase (Ps): Ps-atzF
biuret hydrolase (Ps): Ps-atzE
3.5.1.84
H₂O
ammonia
HO
O
H₂N
NH
O
urea-1-carboxylate
allophanate hydrolase (Ps): Ps-atzF
3.5.1.54
2 CO₂
2 ammonia
H₂O

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.

537 Compounds have structure differences between MetaCyc and HpyCyc.

5 Compounds have differences in slot N+1-NAME between MetaCyc and HpyCyc.

5 Compounds have differences in slot N-1-NAME between MetaCyc and HpyCyc.

4 Compounds have differences in slot N-NAME between MetaCyc and HpyCyc.

2 Compounds have differences in slot OVERVIEW-NODE-SHAPE between MetaCyc and HpyCyc.

42 Compounds have differences in slot COMMON-NAME between MetaCyc and HpyCyc.

7 Compounds have differences in slot CITATIONS between MetaCyc and HpyCyc.

2 Compounds have differences in slot GIBBS-0 between MetaCyc and HpyCyc.

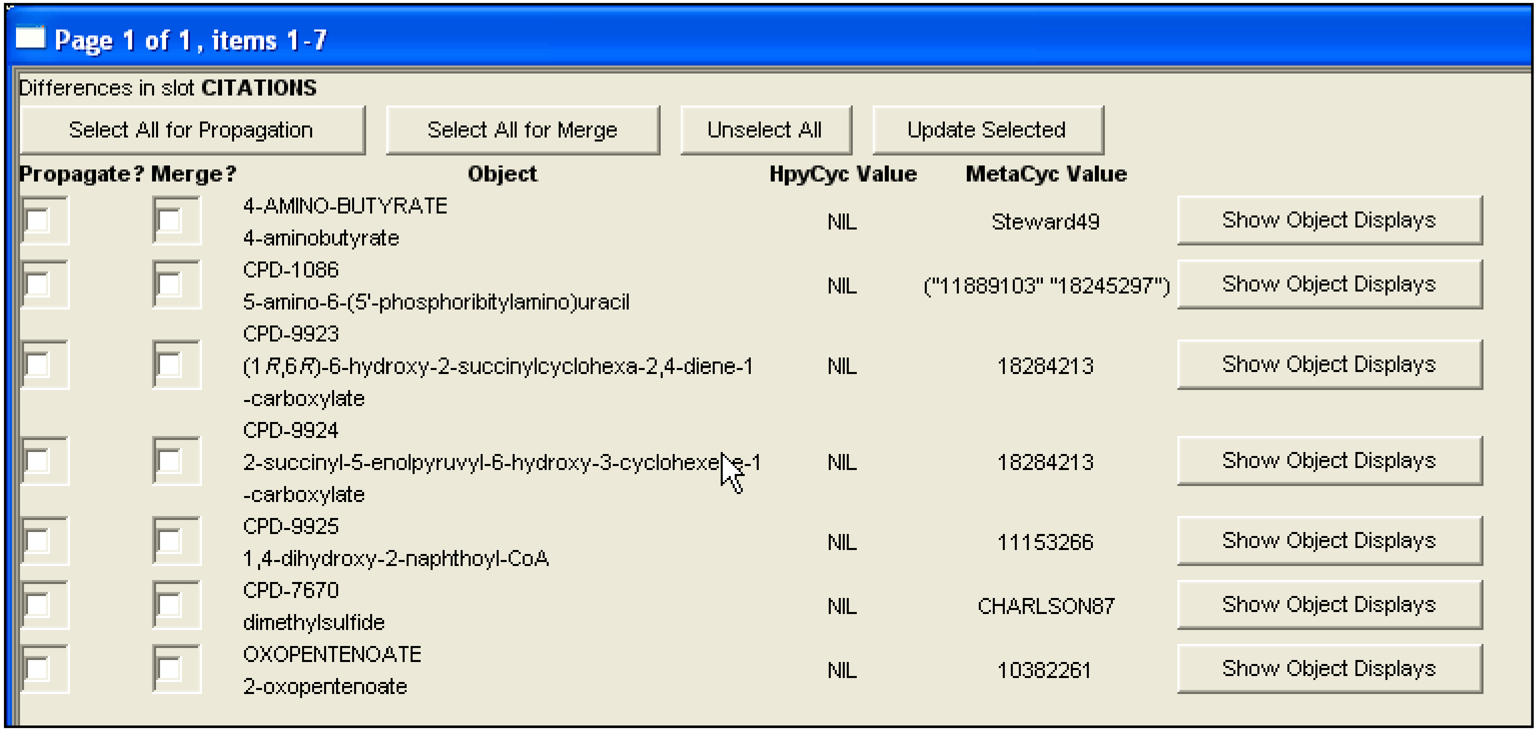
540 Compounds have differences in slot DBLINKS between MetaCyc and HpyCyc.

141 Compounds have differences in slot SYNONYMS between MetaCyc and HpyCyc.

9 Compounds are present in HpyCyc but not in MetaCyc.

Inspecting Differences

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



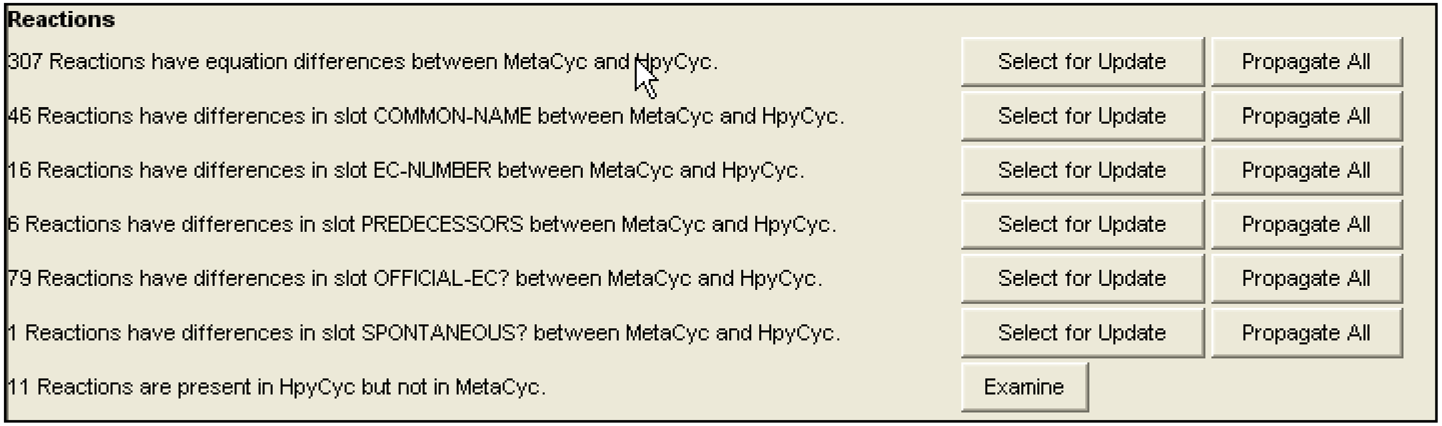
Page 1 of 1, items 1-7

Differences in slot **CITATIONS**

Propagate?	Merge?	Object	HpyCyc Value	MetaCyc Value	
☐	☐	4-AMINO-BUTYRATE 4-aminobutyrate	NIL	Steward49	Show Object Displays
☐	☐	CPD-1086 5-amino-6-(5'-phosphoribitylamino)uracil	NIL	("11889103" "18245297")	Show Object Displays
☐	☐	CPD-9923 (1*R*,6*R*)-6-hydroxy-2-succinylcyclohexa-2,4-diene-1-carboxylate	NIL	18284213	Show Object Displays
☐	☐	CPD-9924 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate	NIL	18284213	Show Object Displays
☐	☐	CPD-9925 1,4-dihydroxy-2-naphthoyl-CoA	NIL	11153266	Show Object Displays
☐	☐	CPD-7670 dimethylsulfide	NIL	CHARLSON87	Show Object Displays
☐	☐	OXOPENTENOATE 2-oxopentenoate	NIL	10382261	Show Object Displays

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



Reactions

307 Reactions have equation differences between MetaCyc and HpyCyc.

46 Reactions have differences in slot COMMON-NAME between MetaCyc and HpyCyc.

16 Reactions have differences in slot EC-NUMBER between MetaCyc and HpyCyc.

6 Reactions have differences in slot PREDECESSORS between MetaCyc and HpyCyc.

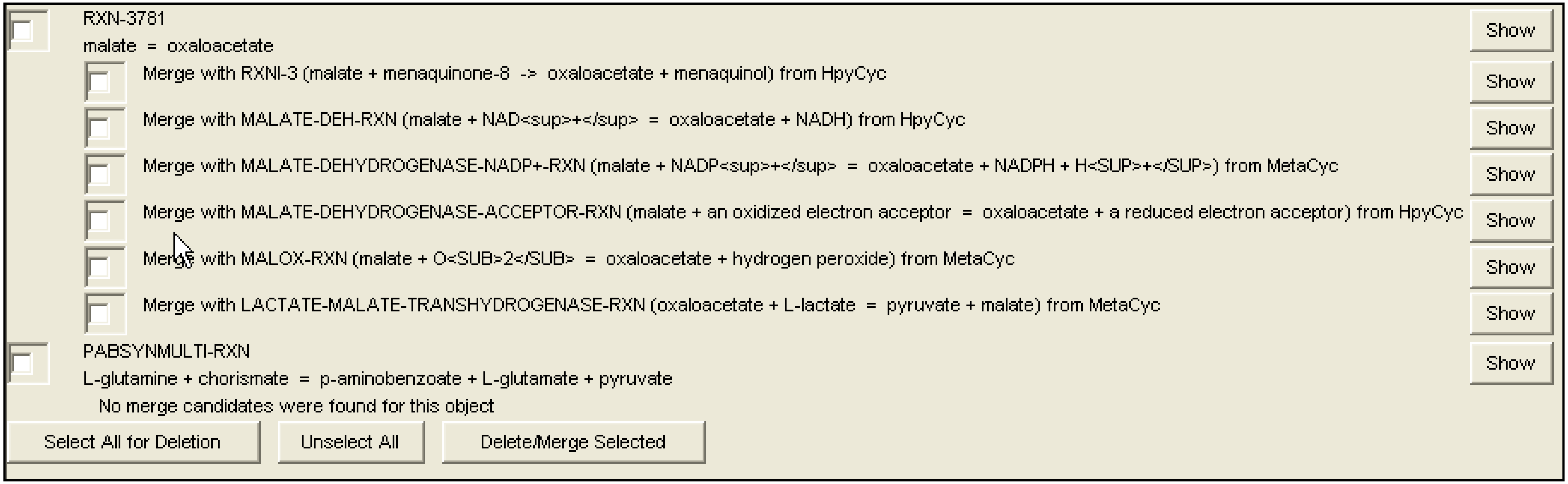
79 Reactions have differences in slot OFFICIAL-EC? between MetaCyc and HpyCyc.

1 Reactions have differences in slot SPONTANEOUS? between MetaCyc and HpyCyc.

11 Reactions are present in HpyCyc but not in MetaCyc.

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



☐ RXN-3781
malate = oxaloacetate

☐ Merge with RXNI-3 (malate + menaquinone-8 -> oxaloacetate + menaquinol) from HpyCyc

☐ Merge with MALATE-DEH-RXN (malate + NAD⁺ = oxaloacetate + NADH) from HpyCyc

☐ Merge with MALATE-DEHYDROGENASE-NADP+-RXN (malate + NADP⁺ = oxaloacetate + NADPH + H⁺) from MetaCyc

☐ Merge with MALATE-DEHYDROGENASE-ACCEPTOR-RXN (malate + an oxidized electron acceptor = oxaloacetate + a reduced electron acceptor) from HpyCyc

☐ Merge with MALOX-RXN (malate + O₂ = oxaloacetate + hydrogen peroxide) from MetaCyc

☐ Merge with LACTATE-MALATE-TRANSHYDROGENASE-RXN (oxaloacetate + L-lactate = pyruvate + malate) from MetaCyc

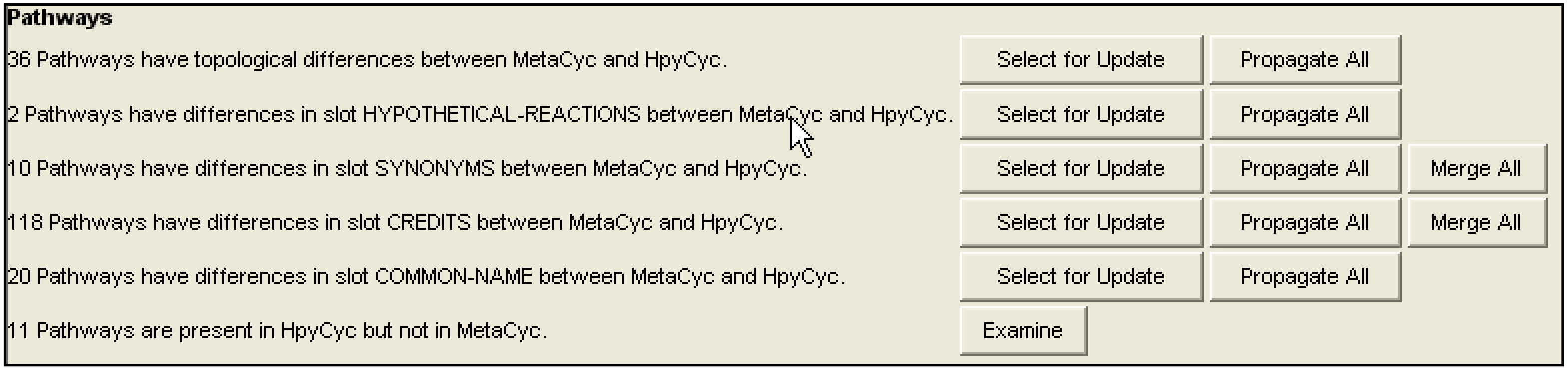
☐ PABSYNMULTI-RXN
L-glutamine + chorismate = p-aminobenzoate + L-glutamate + pyruvate
No merge candidates were found for this object

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!



Pathways

36 Pathways have topological differences between MetaCyc and HpyCyc.

2 Pathways have differences in slot HYPOTHETICAL-REACTIONS between MetaCyc and HpyCyc.

10 Pathways have differences in slot SYNONYMS between MetaCyc and HpyCyc.

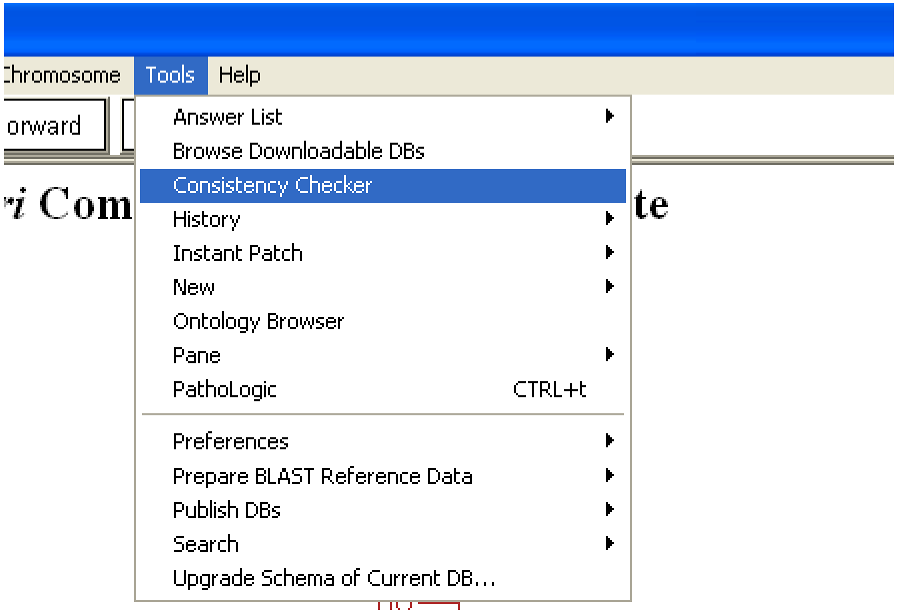
118 Pathways have differences in slot CREDITS between MetaCyc and HpyCyc.

20 Pathways have differences in slot COMMON-NAME between MetaCyc and HpyCyc.

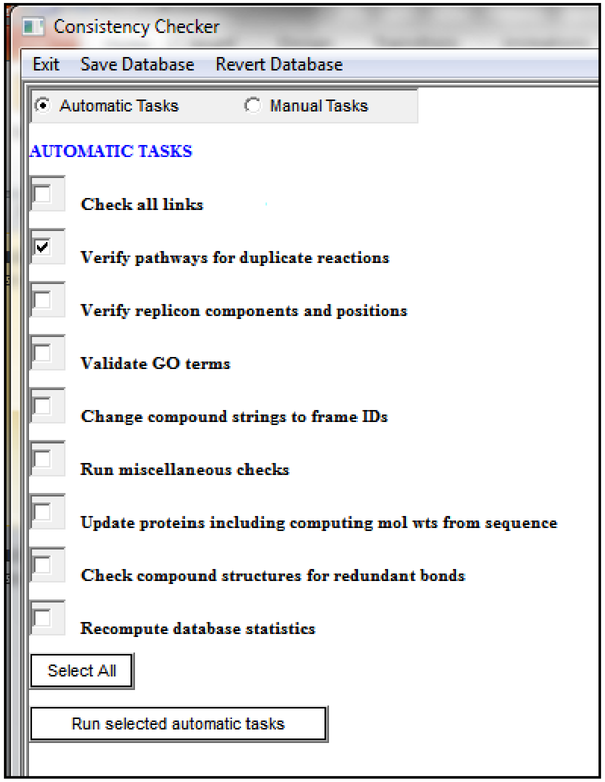
11 Pathways are present in HpyCyc but not in MetaCyc.

The Consistency Checker

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



Automatic and Manual Tasks

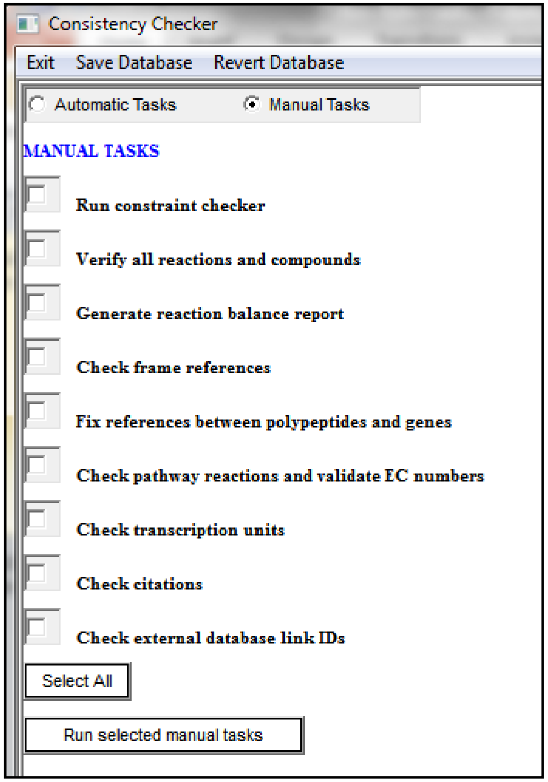


● Automatic Tasks ○ Manual Tasks

AUTOMATIC TASKS

- □ Check all links
- ☑ Verify pathways for duplicate reactions
- □ Verify replicon components and positions
- □ Validate GO terms
- □ Change compound strings to frame IDs
- □ Run miscellaneous checks
- □ Update proteins including computing mol wts from sequence
- □ Check compound structures for redundant bonds
- □ Recompute database statistics

Select All
Run selected automatic tasks


Consistency Checker
Exit Save Database Revert Database

○ Automatic Tasks ● Manual Tasks

MANUAL TASKS

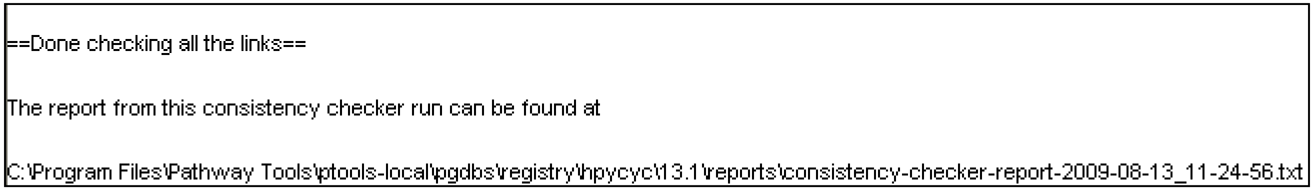
- □ Run constraint checker
- □ Verify all reactions and compounds
- □ Generate reaction balance report
- □ Check frame references
- □ Fix references between polypeptides and genes
- □ Check pathway reactions and validate EC numbers
- □ Check transcription units
- □ Check citations
- □ Check external database link IDs

Select All
Run selected 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.

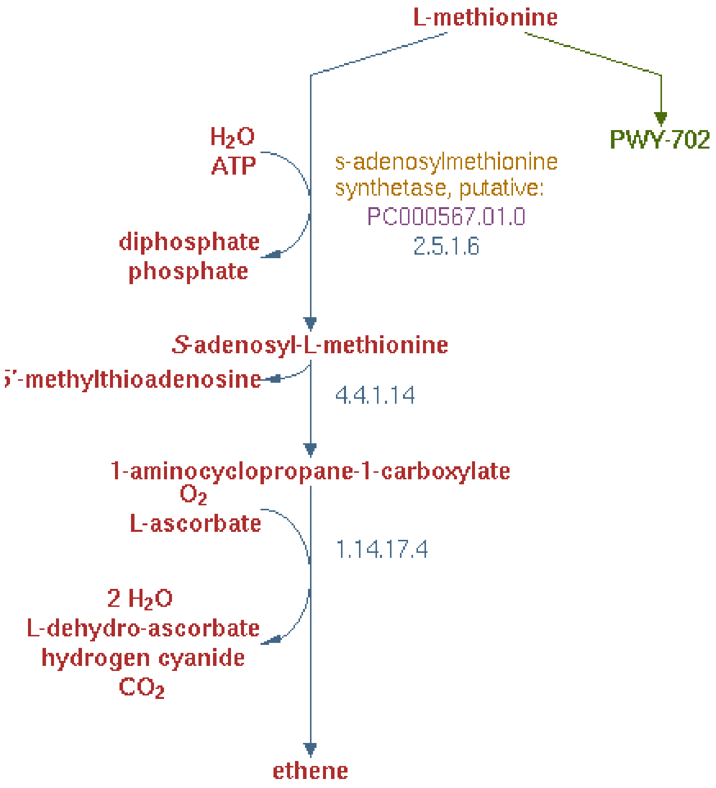


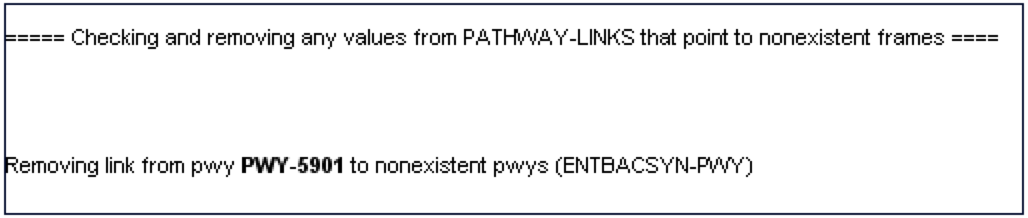
```
==Done checking all the links==

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



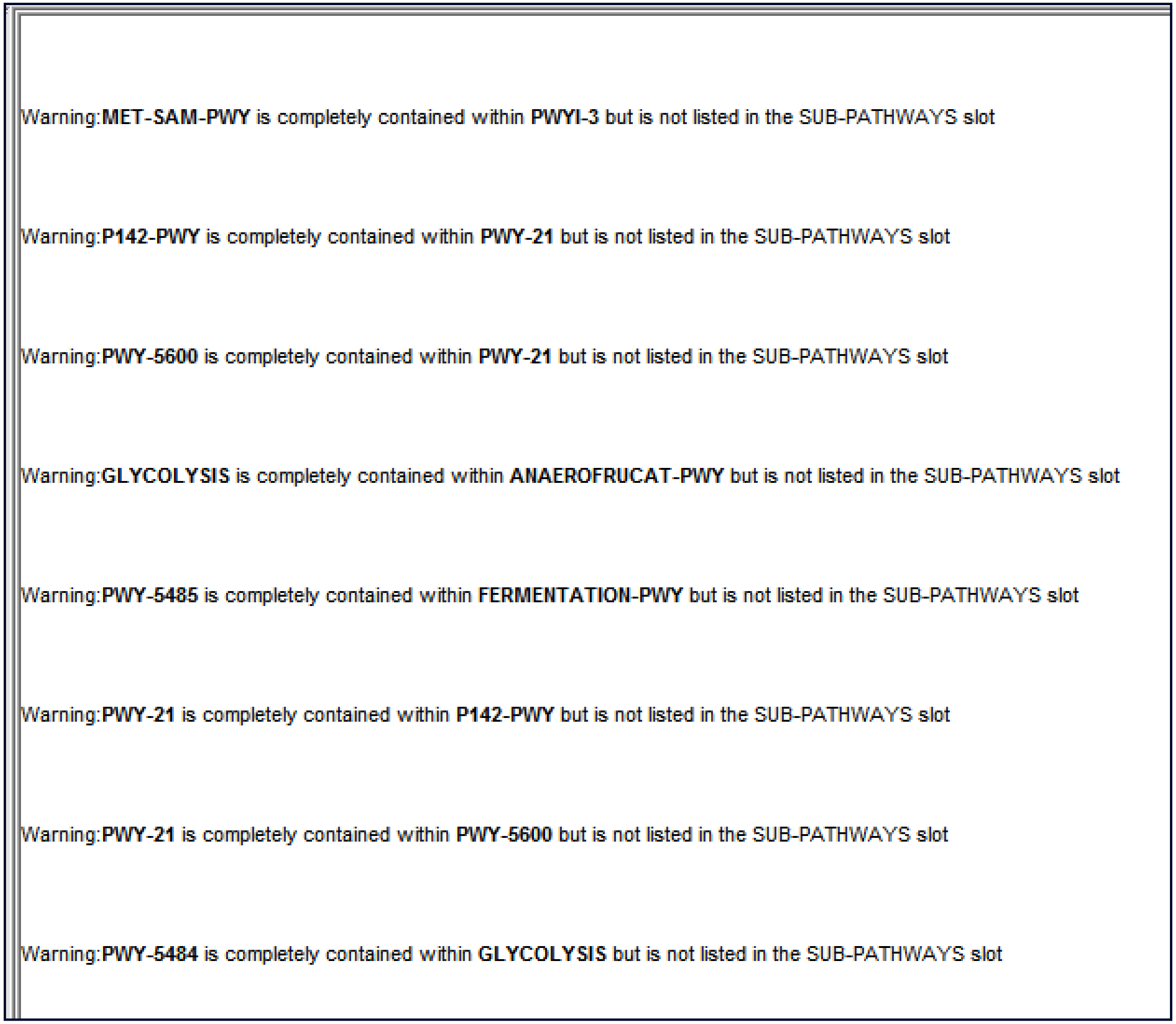


```
===== Checking and removing any values from PATHWAY-LINKS that point to nonexistent frames ====



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.

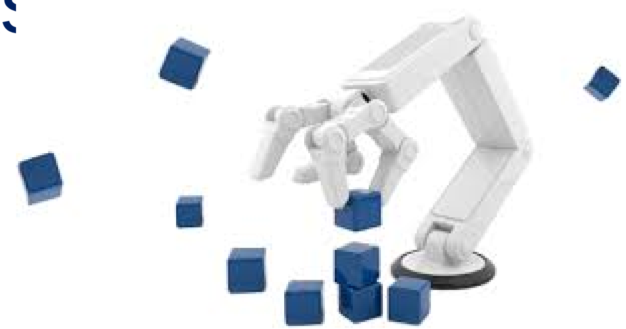


```
Warning:MET-SAM-PWY is completely contained within PWYI-3 but is not listed in the SUB-PATHWAYS slot

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



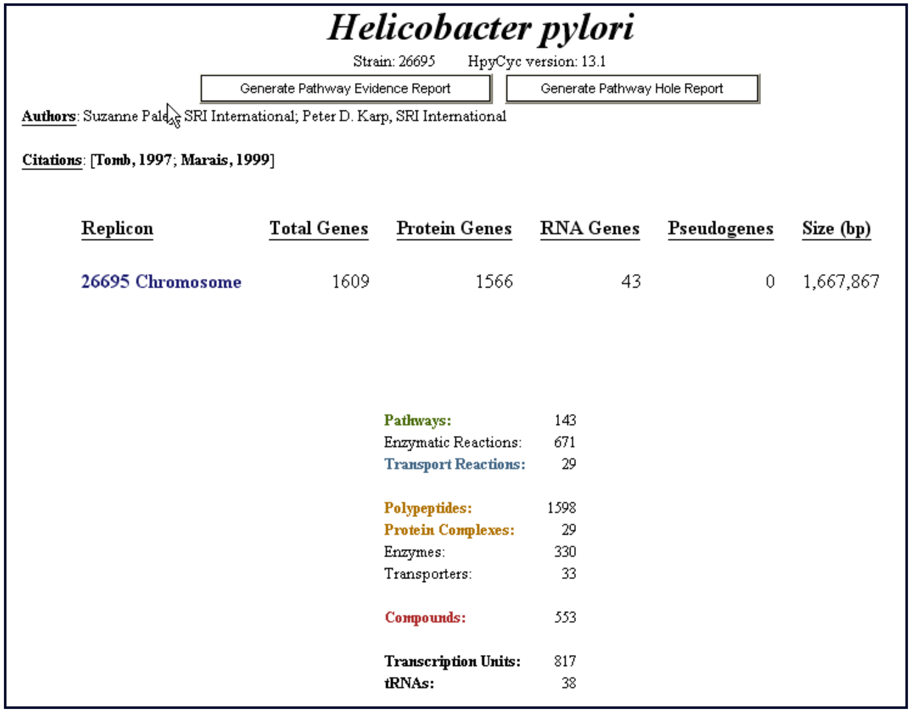
- Run miscellaneous checks: formatting glitches in names, validity of superpathways, clears values of computed slots, deletes temporary frames created by breaks when the pathway editor runs

- 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 Paley, 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.

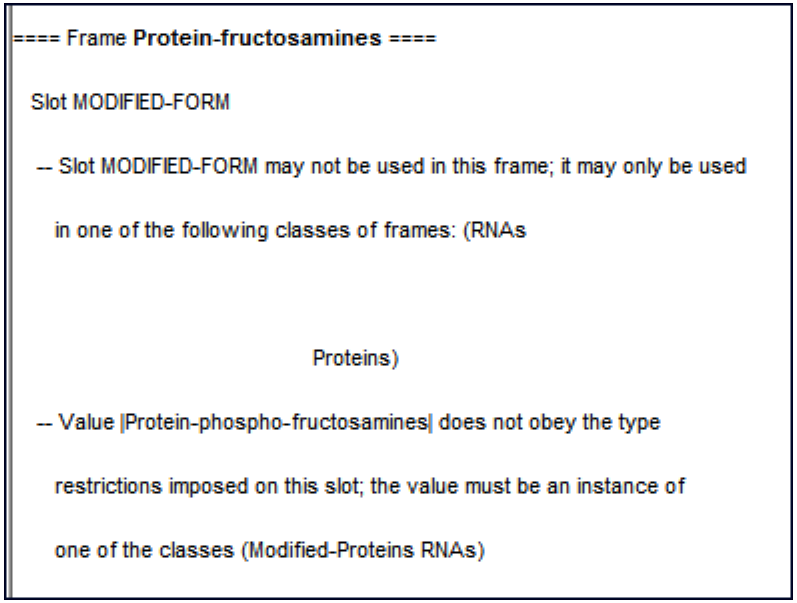
Flags constraints issues. For example, if a slot is supposed to contain only compound frame IDs, but a different type of frame is listed among its values, the constraint checker identifies and flags the offensive value.

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.

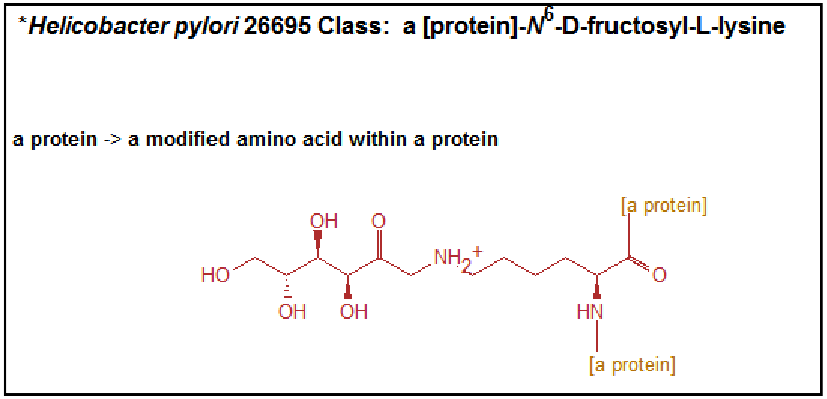
(this means errors are often listed multiple times, under different frames)

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

Constraint Checker Error Reports (example 1)

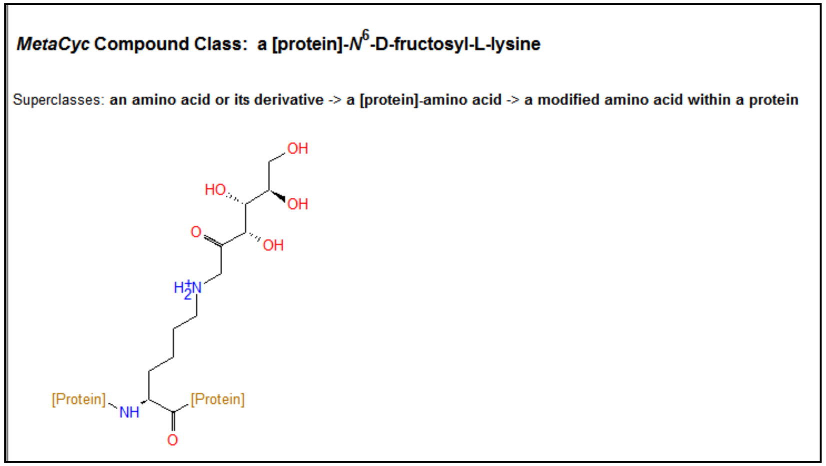


```
==== Frame Protein-fructosamines ====

 Slot MODIFIED-FORM

  -- Slot MODIFIED-FORM may not be used in this frame; it may only be used

     in one of the following classes of frames: (RNAs


                                      Proteins)

  -- Value |Protein-phospho-fructosamines| does not obey the type

     restrictions imposed on this slot; the value must be an instance of

     one of the classes (Modified-Proteins RNAs)
```



****Helicobacter pylori* 26695 Class: a [protein]-*N*^{6}-D-fructosyl-L-lysine**

a protein -> a modified amino acid within a protein



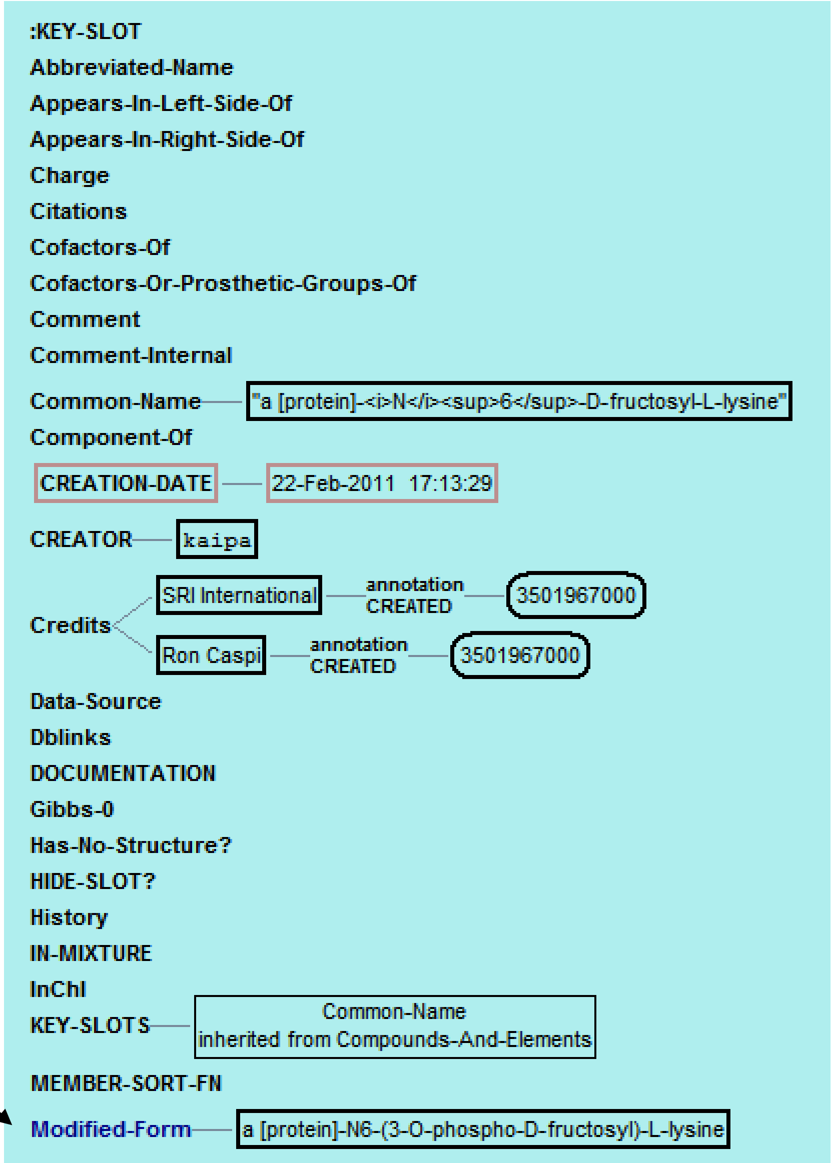
***MetaCyc* Compound Class: a [protein]-*N*^{6}-D-fructosyl-L-lysine**

Superclasses: **an amino acid or its derivative** -> **a [protein]-amino acid** -> **a modified amino acid within a protein**

Obviously, this frame used to be classified as a protein, but has been converted at some point to a chemical compound. Thus, it should no longer contain a Modified-Protein slot.

Fixing The Problem

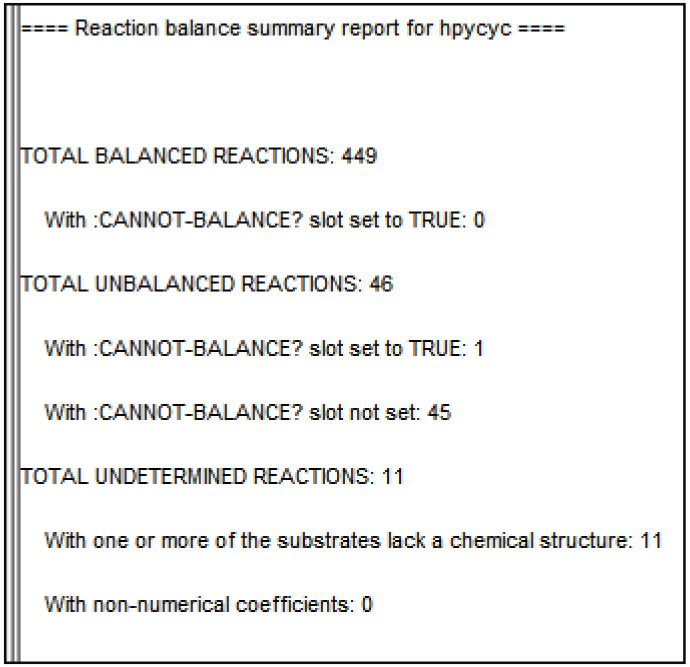
The problematic slot shows up in blue. To solve the problem, highlight the attached value and remove it.



:KEY-SLOT
Abbreviated-Name
Appears-In-Left-Side-Of
Appears-In-Right-Side-Of
Charge
Citations
Cofactors-Of
Cofactors-Or-Prosthetic-Groups-Of
Comment
Comment-Internal
Common-Name — "a [protein]-<i>N</i>⁶-D-fructosyl-L-lysine"
Component-Of
CREATION-DATE — 22-Feb-2011 17:13:29
CREATOR — kaipa
Credits — SRI International — annotation CREATED — 3501967000
Credits — Ron Caspi — annotation CREATED — 3501967000
Data-Source
Dblinks
DOCUMENTATION
Gibbs-0
Has-No-Structure?
HIDE-SLOT?
History
IN-MIXTURE
InChI
KEY-SLOTS — Common-Name inherited from Compounds-And-Elements
MEMBER-SORT-FN
Modified-Form — a [protein]-N6-(3-O-phospho-D-fructosyl)-L-lysine

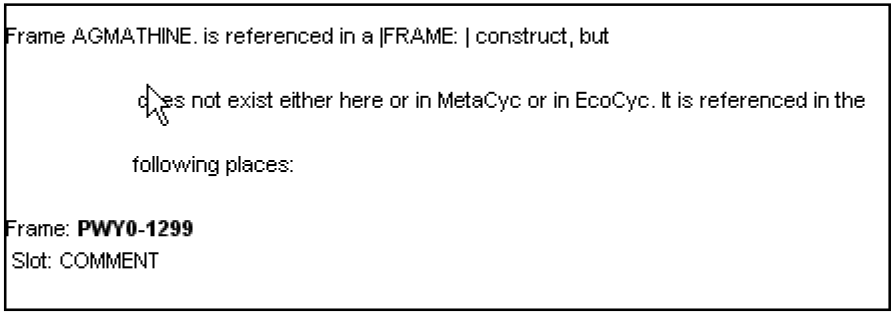
More Manual Tasks

- 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



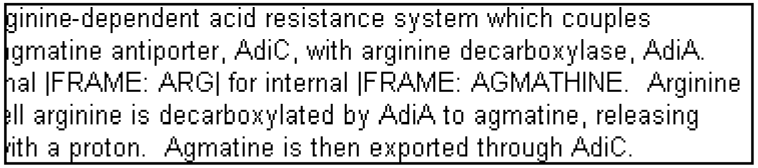
Frame AGMATHINE. is referenced in a |FRAME: | construct, but

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.
hal |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, reports 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!