The consistency Checker, or Overhauling a PGDB

By Ron Caspi

PGDB Atrophy

Your PGDB started out all smooth and shiny…

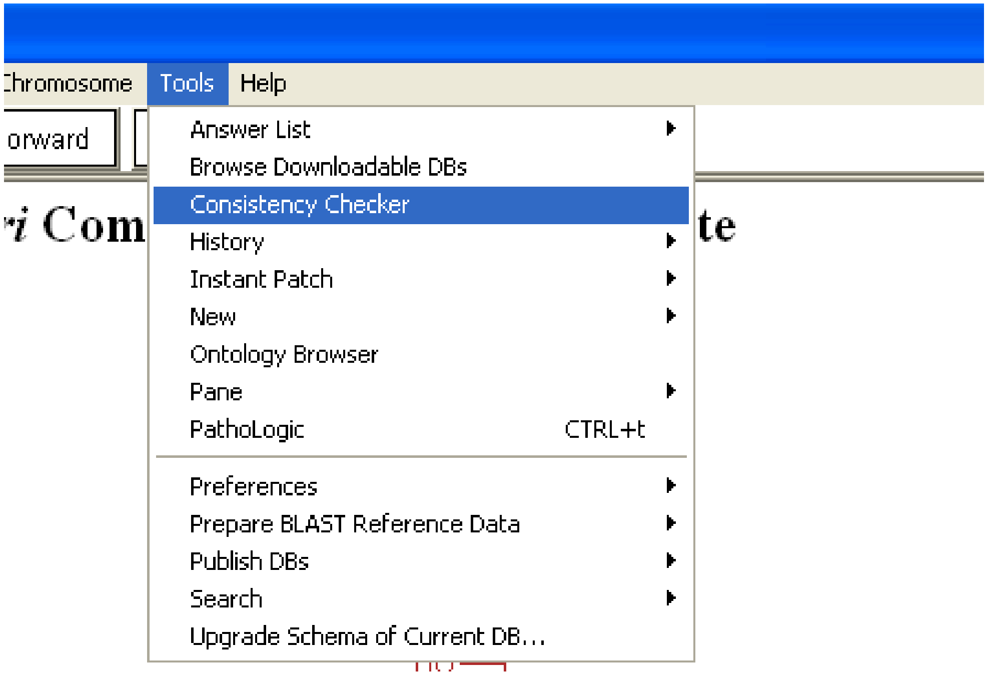
…but after a few years, it looks more like this

It’s time for an overhaul!

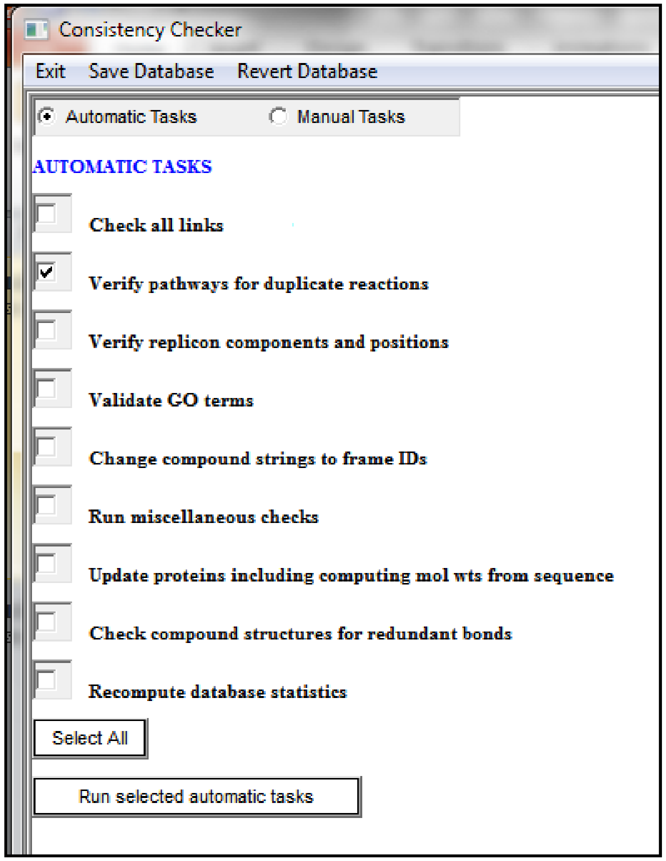
- Update genome annotation
- Propagate updates from Reference DB (MetaCyc)
- Re-run the name matcher
- Rescore pathways
- Re-run the transcription unit predictor
- Run the consistency checker
- Create protein complexes
- Re-run the Transport Inference Parser

The Consistency Checker

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



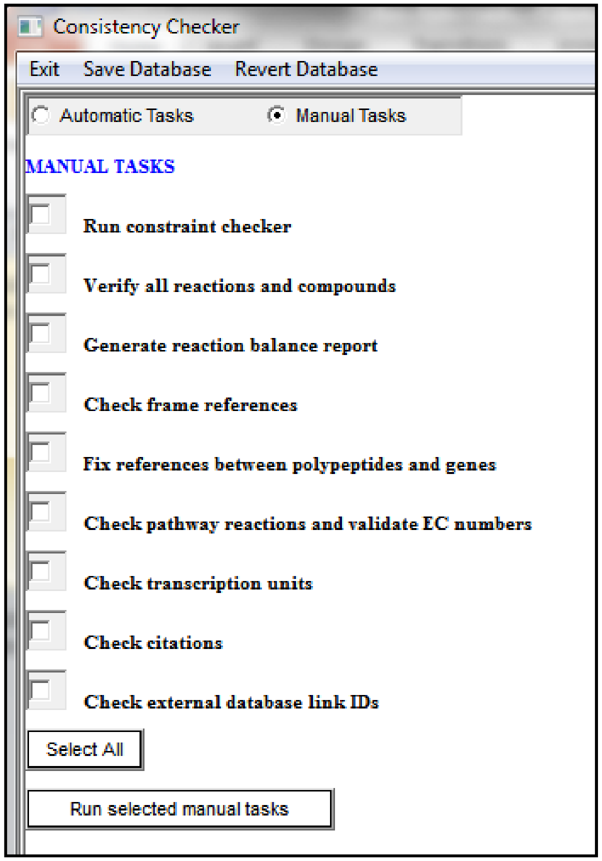
Automatic and Manual Tasks



● 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



○ Automatic Tasks ● Manual Tasks

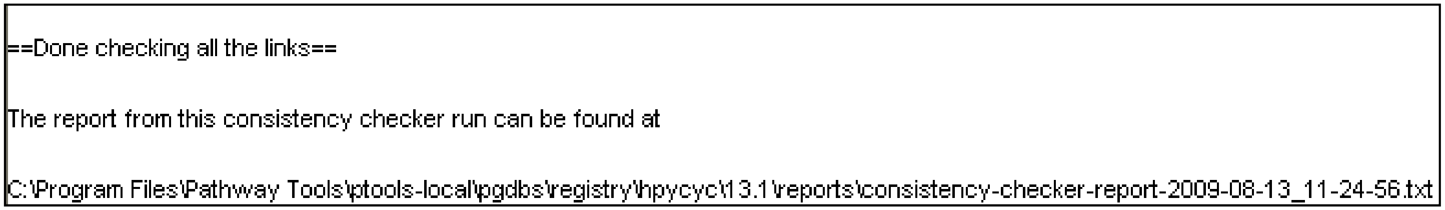
MANUAL TASKS

- □ 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

- 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.

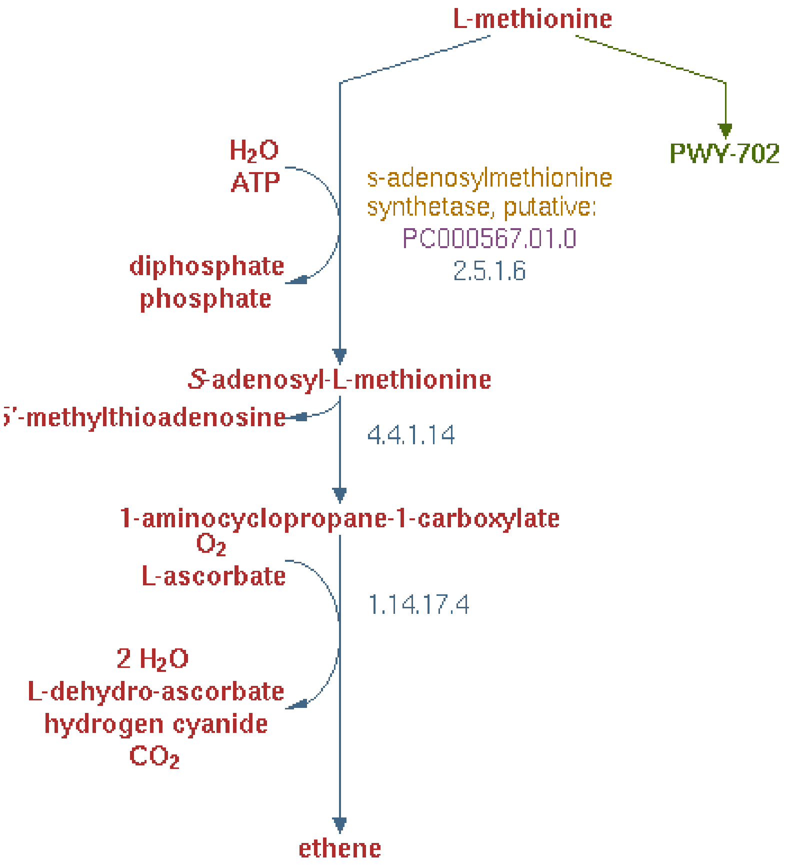


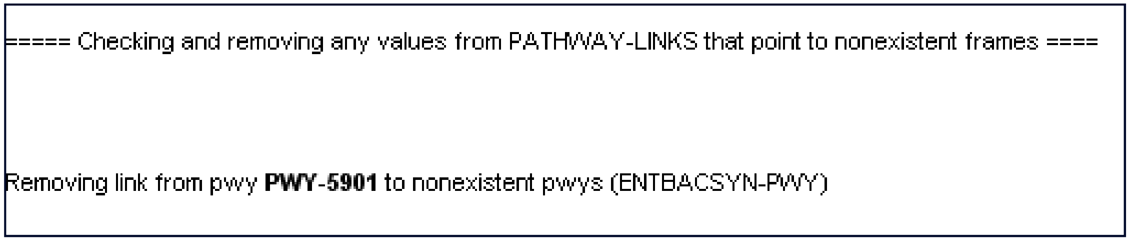
```
==Done checking all the links==

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



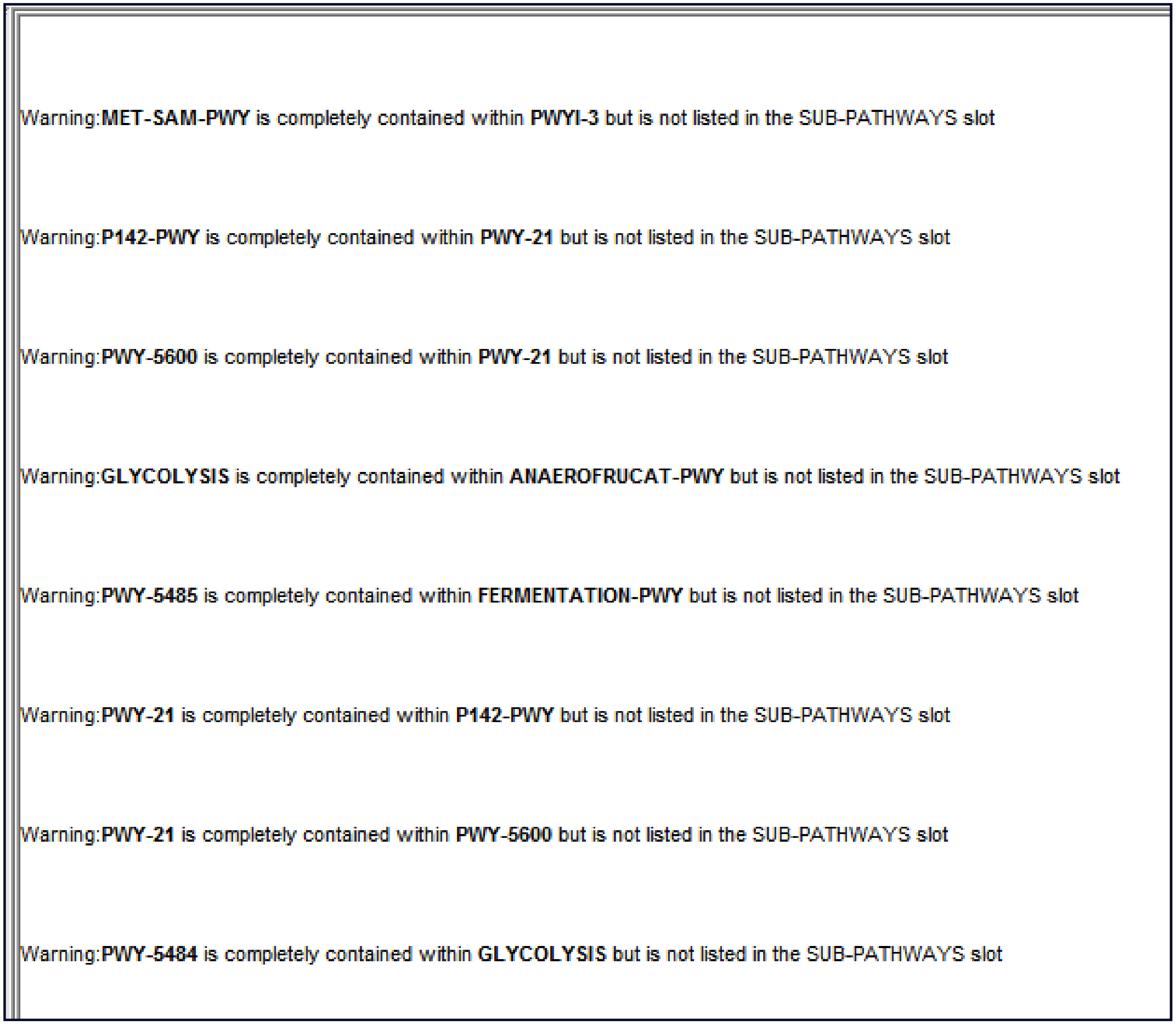


```
===== Checking and removing any values from PATHWAY-LINKS that point to nonexistent frames ====



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 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, 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

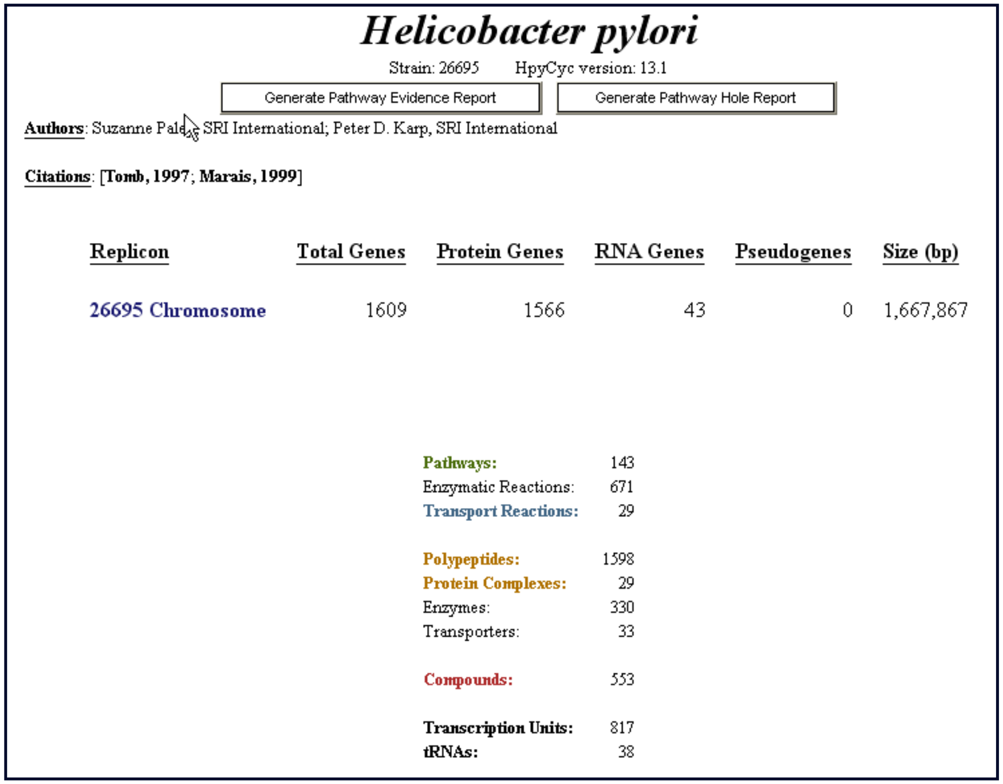
- Run miscellaneous checks: formatting glitches in names, validity of superpathways, clears values of computed slots, deletes temporary frames created by the pathway editor

- Update proteins: molecular weights recalculated from sequence

- Check compound structures for redundant bonds

Automatic Tasks: Recompute database statistics

Its the only way to change 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: Run Constraint Checker

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

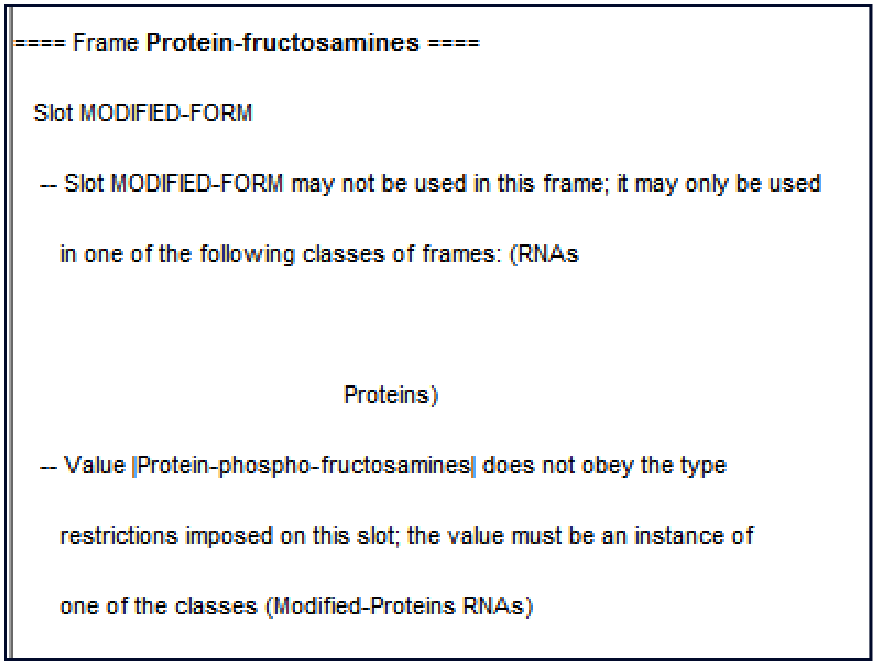
Flags constraints issues. For example, if a slot is supposed to contain only compound frames, but a different type of frame is listed among its values, the constraint checker identifies and flags the offensive value.

The opposite is true as well: the checker will flag that compound as present in a slot of a frame that is not suppose 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.

Run Constraint Checker Error Reports: Example 1

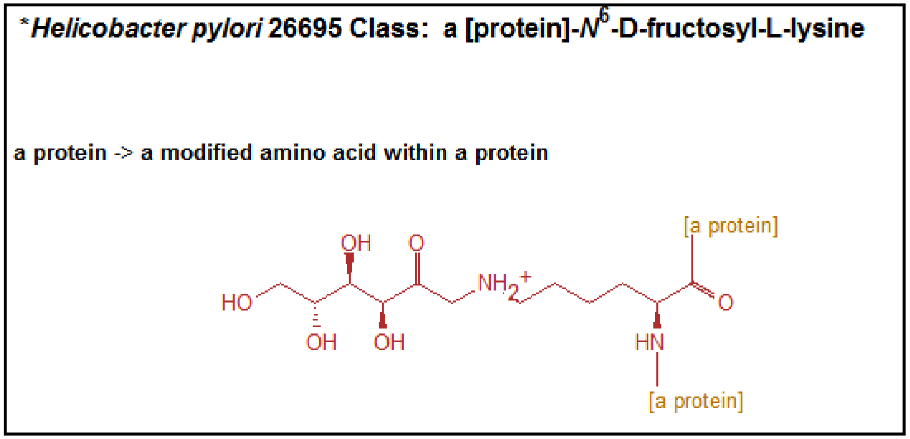


==== Frame **Protein-fructosamines** ====

Slot MODIFIED-FORM

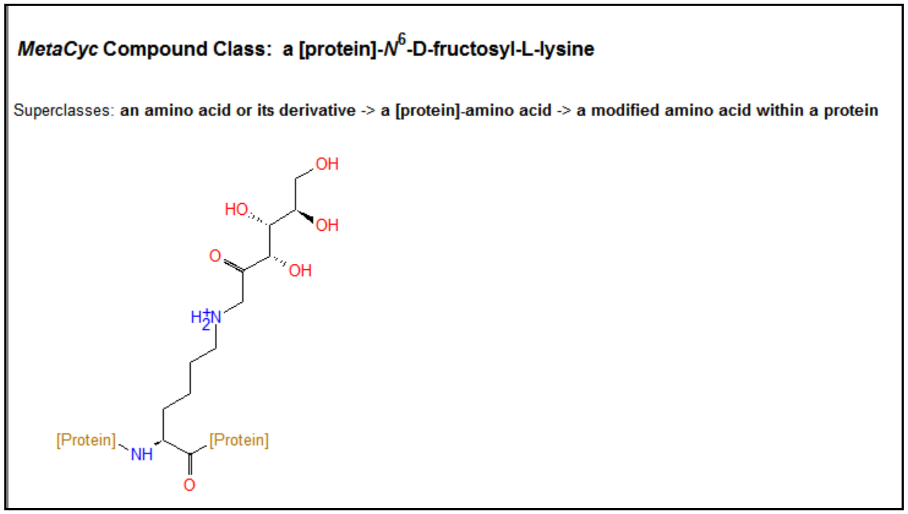
-- Slot MODIFIED-FORM may not be used in this frame; it may only be used in one of the following classes of frames: (RNAs Proteins)

-- 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



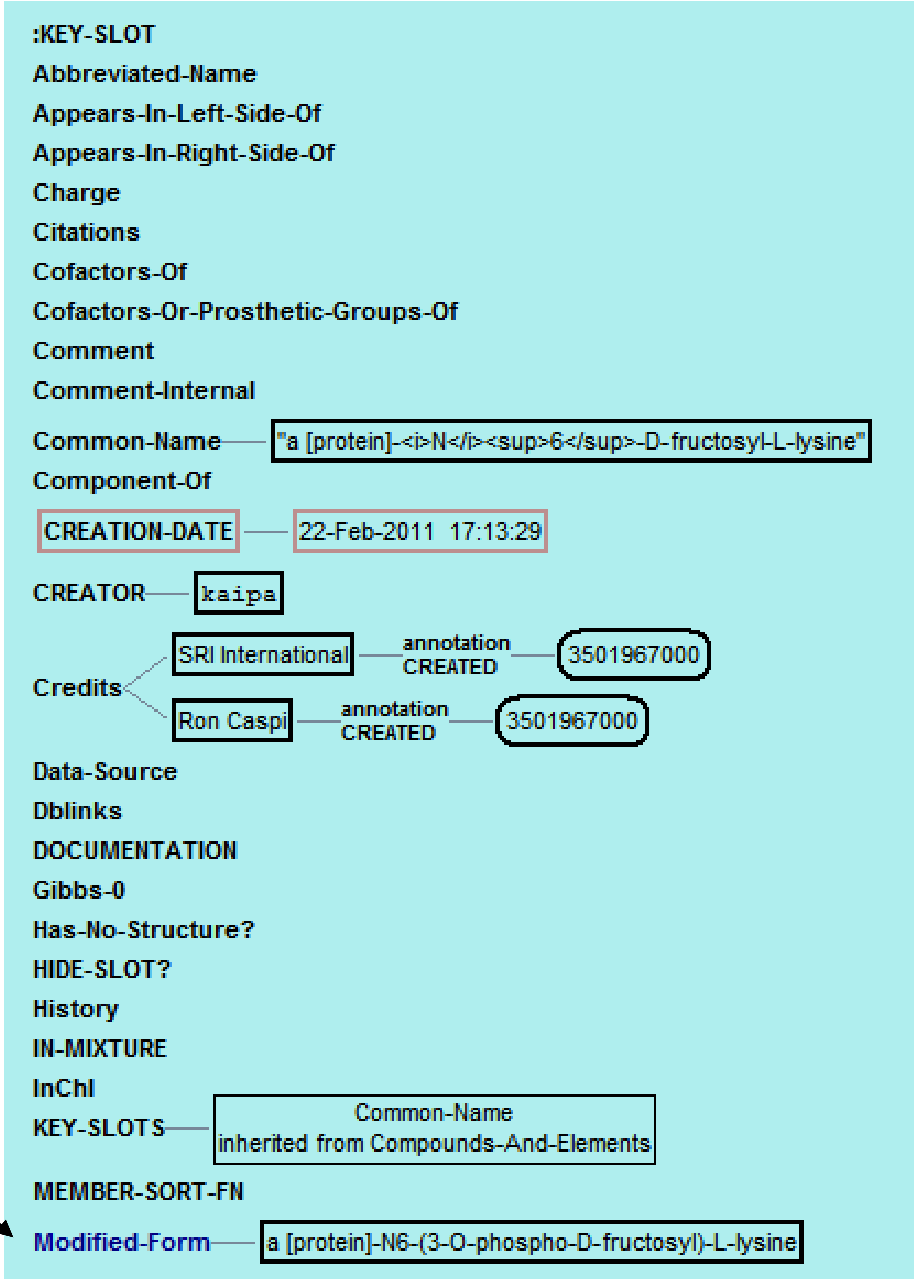
MetaCyc **Compound Class: a [protein]-*N*^{6}-D-fructosyl-L-lysine**

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.

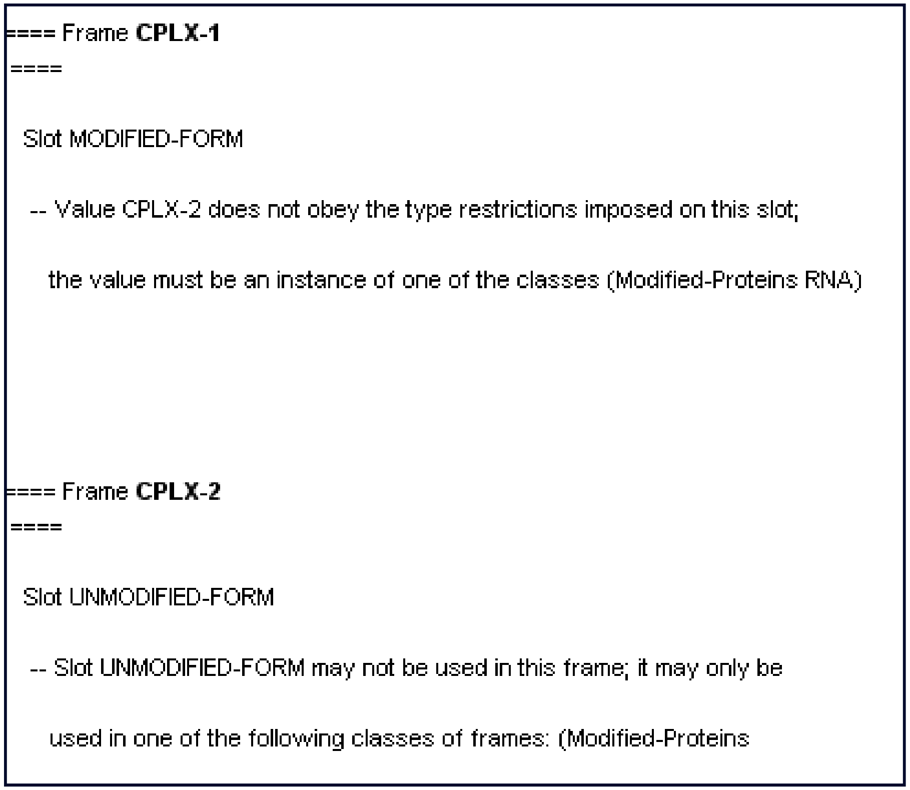
Fixing The Problem

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

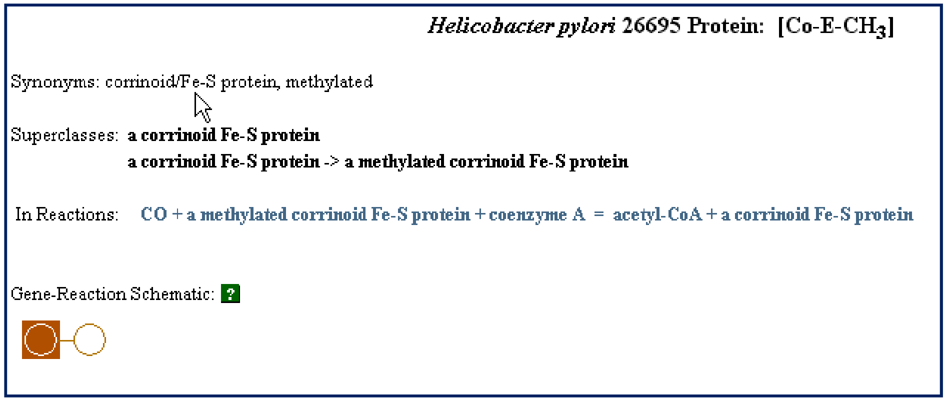


:KEY-SLOT
Abbreviated-Name
Appears-In-Left-Side-Of
Appears-In-Right-Side-Of
Charge
Citations
Cofactors-Of
Cofactors-Or-Prosthetic-Groups-Of
Comment
Comment-Internal
Common-Name — "a [protein]-<i>N</i>⁶-D-fructosyl-L-lysine"
Component-Of
CREATION-DATE — 22-Feb-2011 17:13:29
CREATOR — kaipa
Credits — SRI 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

Constraint Error Reports: Example 2



```
==== Frame CPLX-1
====

  Slot MODIFIED-FORM

   -- Value CPLX-2 does not obey the type restrictions imposed on this slot;

     the value must be an instance of one of the classes (Modified-Proteins RNA)




==== Frame CPLX-2
====

  Slot UNMODIFIED-FORM

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

     used in one of the following classes of frames: (Modified-Proteins
```



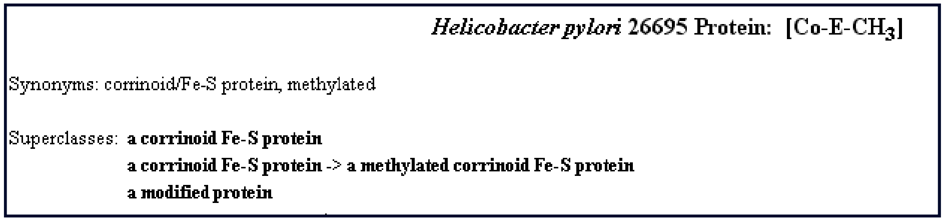
Helicobacter pylori 26695 Protein: [Co-E-CH$_3$]

Synonyms: corrinoid/Fe-S protein, methylated

Superclasses: **a corrinoid Fe-S protein**
a corrinoid Fe-S protein -> a methylated corrinoid Fe-S protein

In Reactions: **CO + a methylated corrinoid Fe-S protein + coenzyme A = acetyl-CoA + a corrinoid Fe-S protein**

Gene-Reaction Schematic: ?



Helicobacter pylori 26695 Protein: [Co-E-CH$_3$]

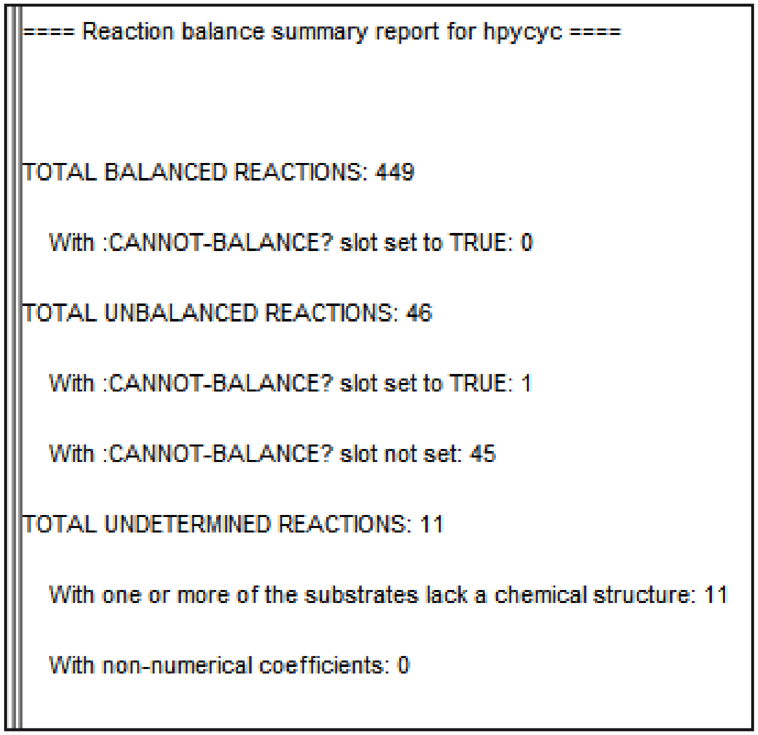
Synonyms: corrinoid/Fe-S protein, methylated

Superclasses: **a corrinoid Fe-S protein**
a corrinoid Fe-S protein -> a methylated corrinoid Fe-S protein
a modified protein

The problem here is that CPLX-2, a modified form of CPLX-1, has not been classified as a modified protein. The solution is to open CPLX-2 in the Protein Editor and classify it as a modified protein.

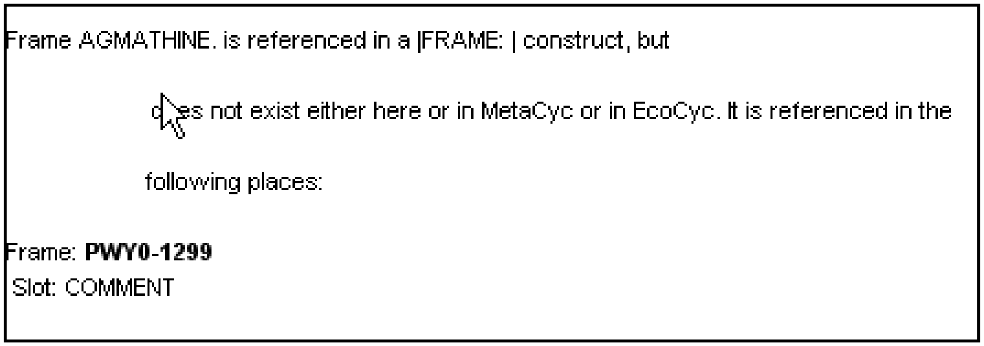
More Manual Tasks

- Verify all reactions and compounds: finds defective 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 Error Report Example



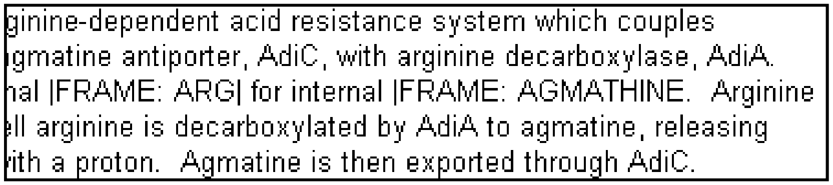
Frame AGMATHINE. is referenced in a |FRAME: | construct, but

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, and scans it 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.

And When You Finish, take pride at your newly renovated PGDB!