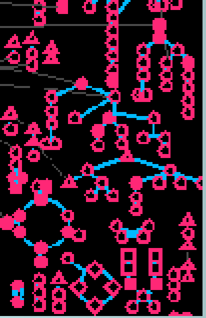
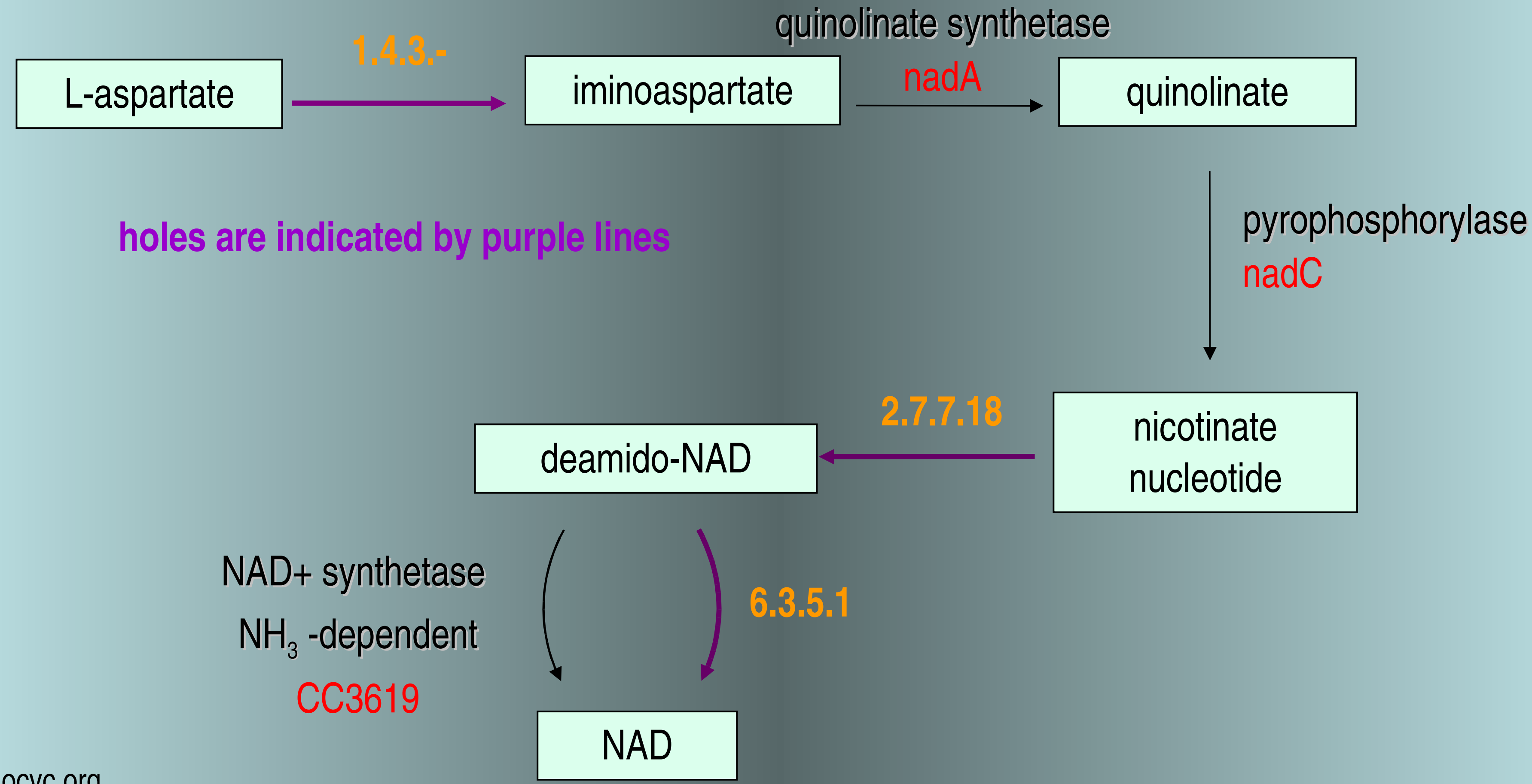


Identify Pathway Hole Fillers

Definition: Pathway Holes are reactions in metabolic pathways for which no enzyme is identified in the PGDB.



L-aspartate

1.4.3.-

iminoaspartate

quinolinate synthetase

nadA

quinolinate

pyrophosphorylase

nadC

nicotinate nucleotide

2.7.7.18

deamido-NAD

NAD+ synthetase

NH_3 -dependent

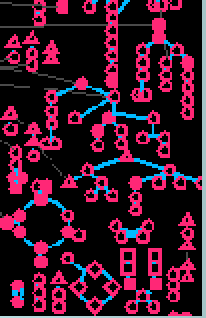
CC3619

6.3.5.1

NAD

holes are indicated by purple lines

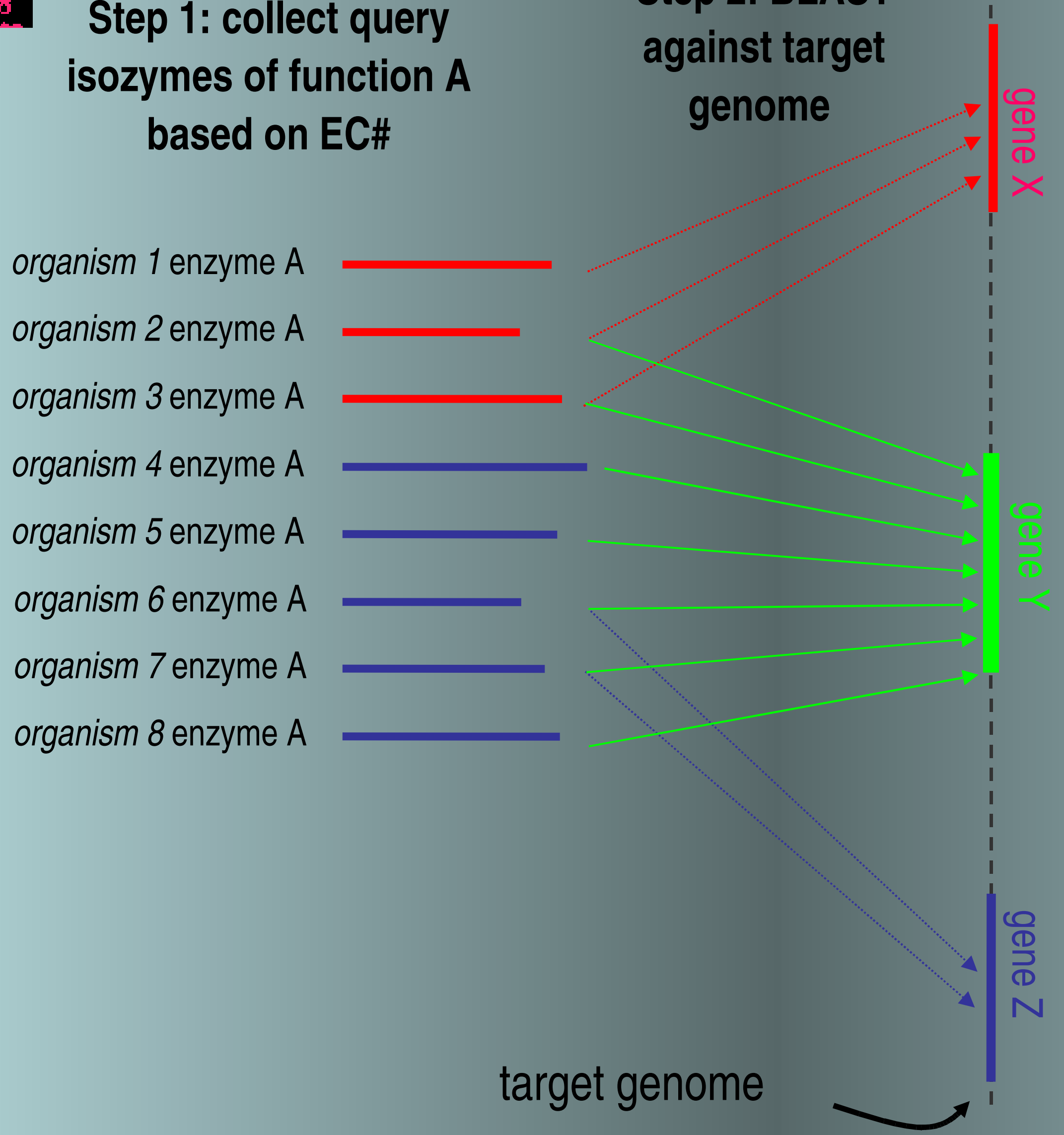
Biocyc.org



Algorithm for identifying candidates and consolidating data…

Step 1: collect query isozymes of function A based on EC#

Step 2: BLAST against target genome



Step 3 & 4: Consolidate hits and evaluate evidence

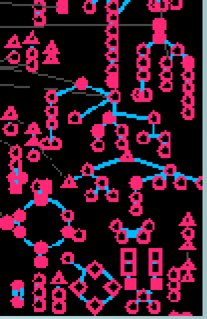
Candidates

Gene X

Gene Y

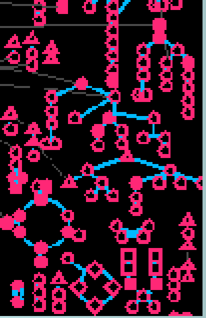
Gene Z

Biocyc.org

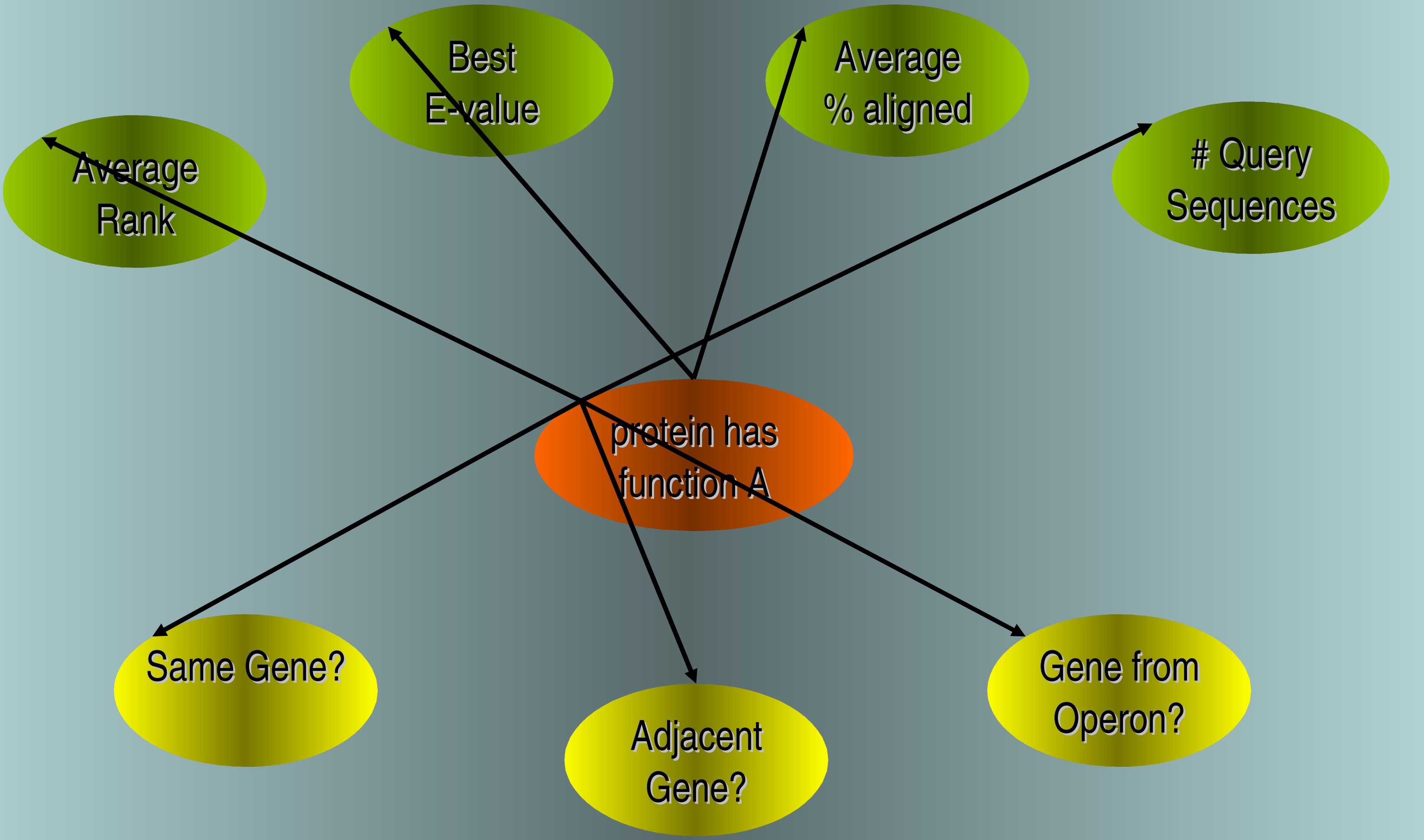


Features used to calculate the probability that a protein has the desired function…

- Best E-value
- Avg. rank
- Avg % aligned
- Number of query sequences aligned
- Candidate in same directon as another pathway gene?
- Candidate is adjacent to a gene that catalyzes an adjacent reaction?
- Candidate catalyzes another pathway reaction?



Use Bayesian classifier to evaluate candidates



Best E-value

Average % aligned

Average Rank

# Query Sequences

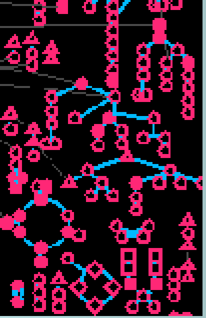
protein has function A

Same Gene?

Adjacent Gene?

Gene from Operon?

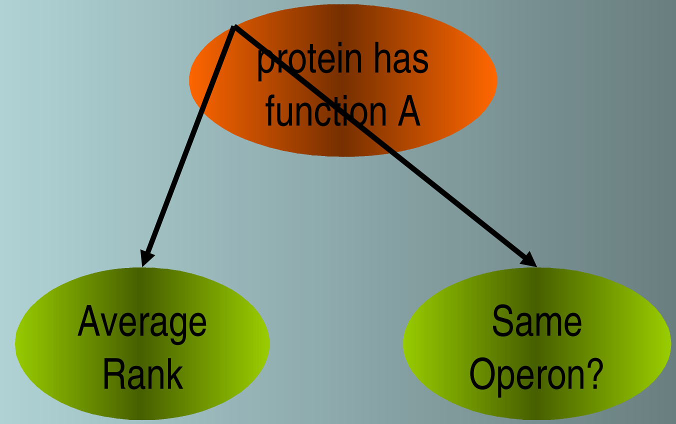
Biocyc.org



Computing P(has function)

Apply Bayes' rule:

$$P(true|evidence)=\frac{P(true)P(evidence|true)}{P(true)P(evidence|true)+P(false)P(evidence|false)}$$



protein has function A

Average Rank

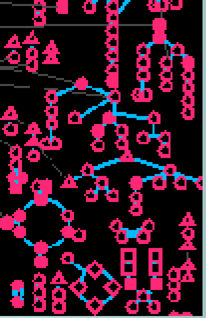
Same Operon?

Compute probability distributions, i.e., P(evidence|true) and P(evidence|false), from the "known" reactions in the database.

e.g., Same operon?

In operon?	True hit Has-Fn(A)	False hit ~Has-Fn(A)
yes	0.24 (TP)	0.04 (FP)
no	0.76 (FN)	0.96 (TN)

Biocyc.org

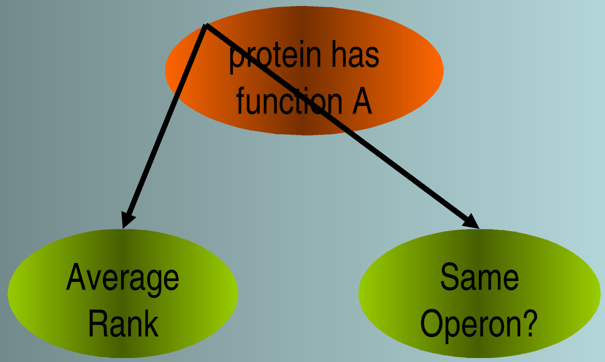


Computing P(has function)

Example:

Candidate X has avg-rank 1.5 and is in a directon with another pathway gene.

From training data:



P(average-rank = 1.5 | has-function) = 0.40

P(average-rank = 1.5 | ¬has-function) = 0.03

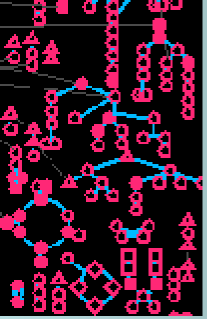
P(pathway-directon = true | has-function) = 0.24

P(pathway-directon = true | ¬has-function) = 0.04

P(has-function) = 0.041 (4.1% of candidates in training data are true hits)

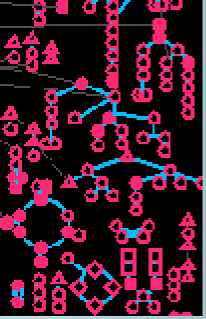
$$P(has_{function_A}) = \frac{0.041 * 0.40 * 0.24}{0.041 * 0.40 * 0.24 + 0.959 * 0.03 * 0.04}$$

$$0.77$$



Steps that must be completed <u>before</u> running the Pathway Hole Filler

- Install BLAST executable (should already be installed on training room machines)
- Prepare BLAST protein db
  - Need FASTA format genome nucleotide sequence (see me if you have something different, like ESTs, or have no nucleotide sequence data file)
- In general, the more pathways in your PGDB, the more candidates the pathway hole filler will have to find

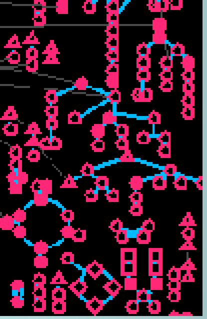


Conceptual stages of the pathway hole filler

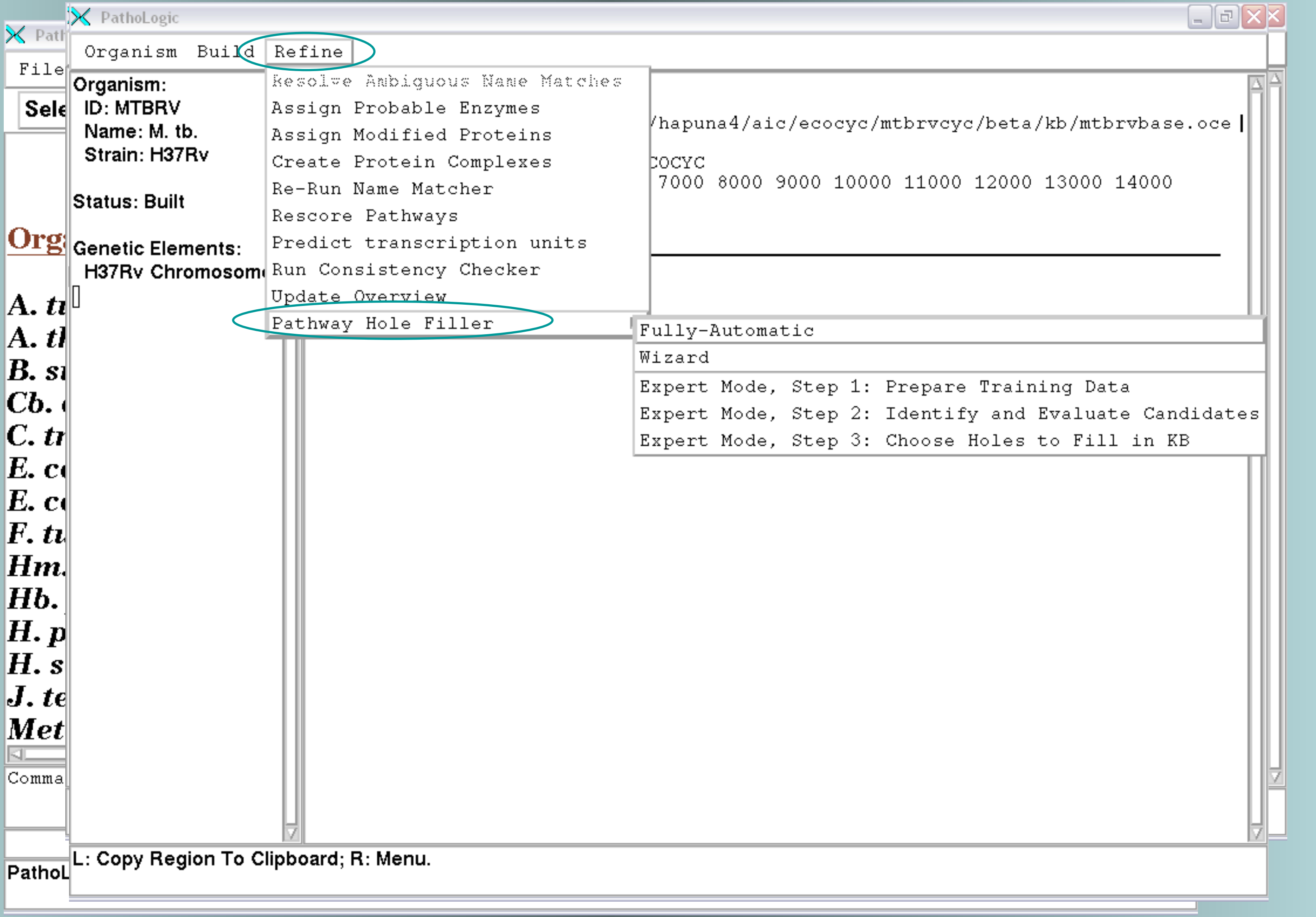
1. Prepare training data for Bayes classifier
   - Collect feature data for known rxns in PGDB
   - Calculate probability distributions for classifier

2. Identify and evaluate candidates
   - Collect feature data for each candidate
   - Use classifier to determine P(has-function)

3. Choose holes to fill in KB
   - Either select all above a cut-off or manually review candidates



Navigating to the Pathway Hole Filler



Organism:
ID: MTBRV
Name: M. tb.
Strain: H37Rv

Status: Built

Genetic Elements:
H37Rv Chromosom

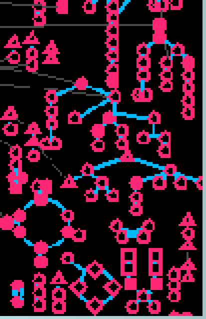
/hapuna4/aic/ecocyc/mtbrvcyc/beta/kb/mtbrvbase.oce

COCYC
7000 8000 9000 10000 11000 12000 13000 14000

Org

A. tu
A. th
B. su
Cb.
C. tr
E. co
E. co
F. tu
Hm.
Hb.
H. p
H. s
J. te
Met

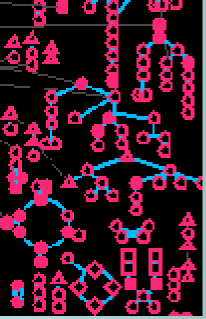
L: Copy Region To Clipboard; R: Menu.



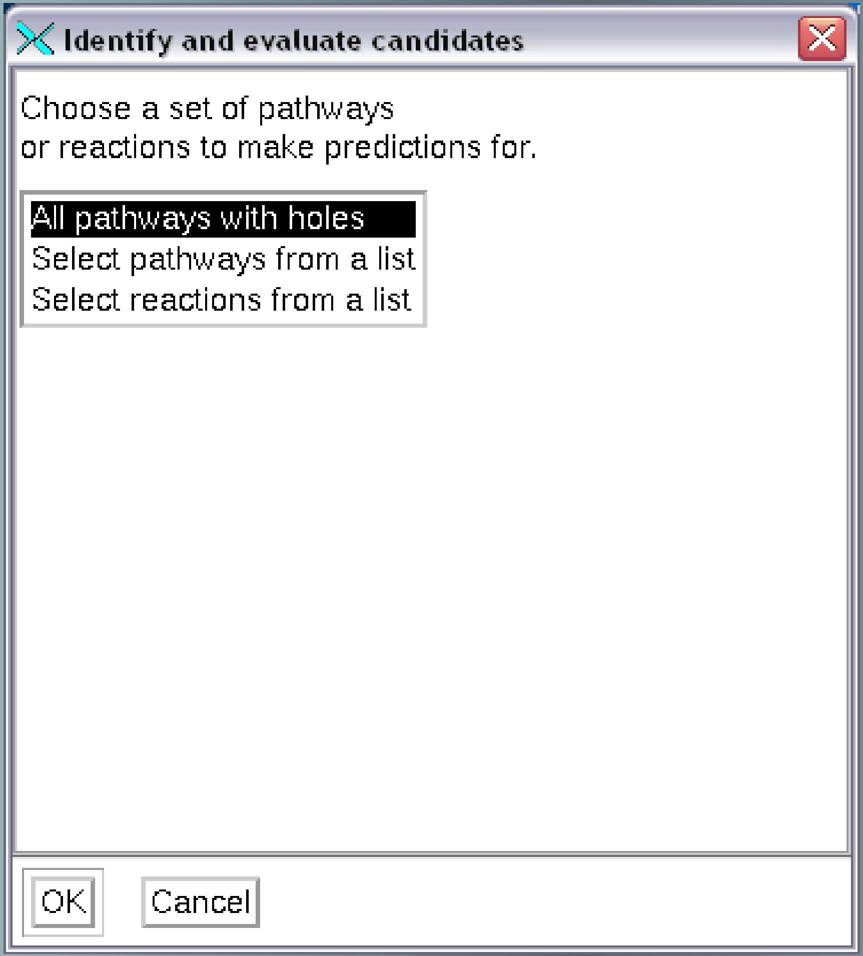
Step 1: Prepare Training Data…

Calculate training data from your organism or use existing training data…

- Once Step 1 has been completed, the training data are saved and can be reused (even in another Pathway Tools session).
- If using existing data from *E. coli* the training data are based on data from the literature.

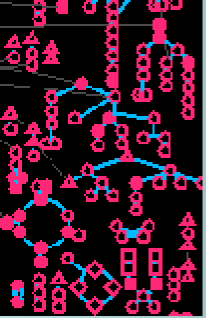


Step 2: Identify & Evaluate Candidates…



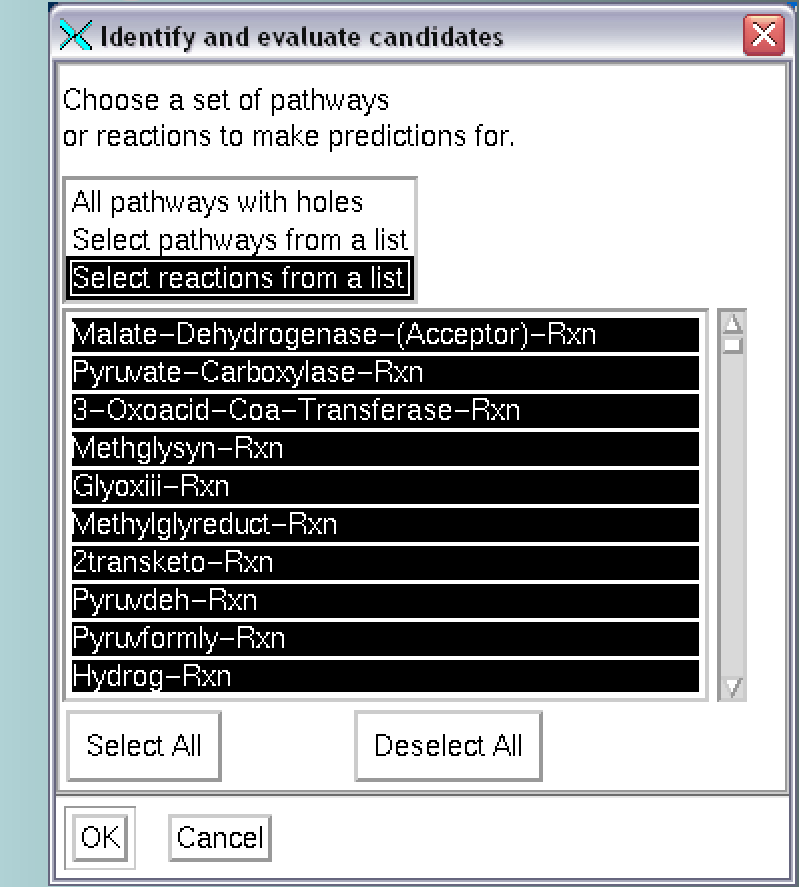
Identify and evaluate candidates

Choose a set of pathways
or reactions to make predictions for.



Step 2: Identify & Evaluate Candidates

Select reactions from a list



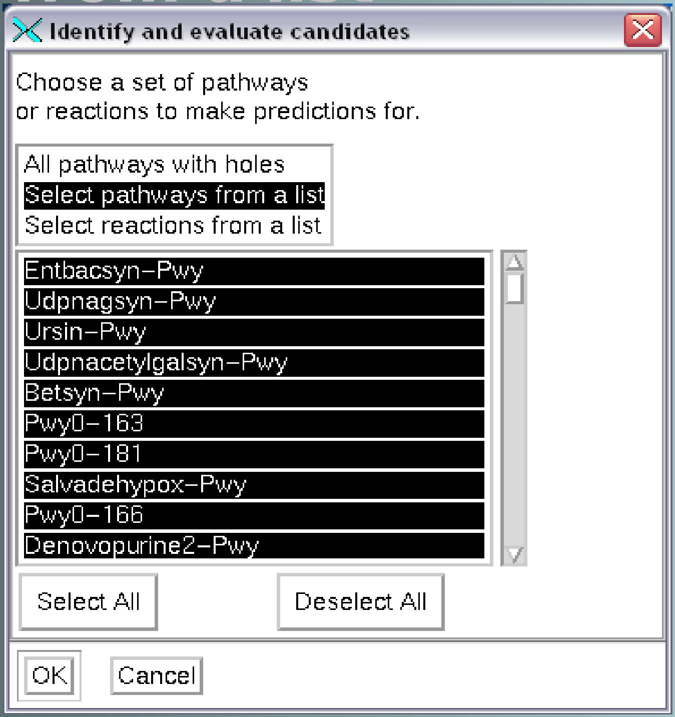
Identify and evaluate candidates

Choose a set of pathways
or reactions to make predictions for.

All pathways with holes
Select pathways from a list
Select reactions from a list

Malate–Dehydrogenase–(Acceptor)–Rxn
Pyruvate–Carboxylase–Rxn
3–Oxoacid–Coa–Transferase–Rxn
Methglysyn–Rxn
Glyoxiii–Rxn
Methylglyreduct–Rxn
2transketo–Rxn
Pyruvdeh–Rxn
Pyruvformly–Rxn
Hydrog–Rxn

Select pathways from a list



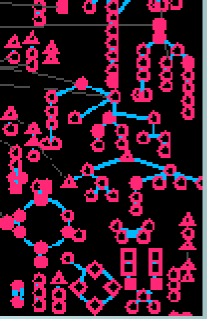
Identify and evaluate candidates

Choose a set of pathways
or reactions to make predictions for.

All pathways with holes
Select pathways from a list
Select reactions from a list

Entbacsyn–Pwy
Udpnagsyn–Pwy
Ursin–Pwy
Udpnacetylgalsyn–Pwy
Betsyn–Pwy
Pwy0–163
Pwy0–181
Salvadehypox–Pwy
Pwy0–166
Denovopurine2–Pwy

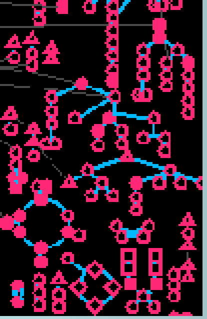
Biocyc.org



Modes of operation…

Fully automatic

- No interaction required from user
- All default values used
  - – Prepare training data – all known rxns in KB
  - – Identify and evaluate candidates – all pathways with pathway holes
  - – Choose holes to fill in KB – all holes with P>0.9 filled
- Evidence code: “Automatic inference from sequence similarity”



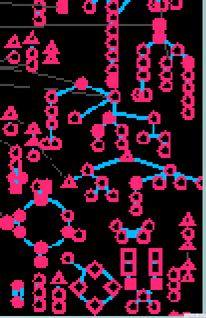
Modes of operation…

Wizard

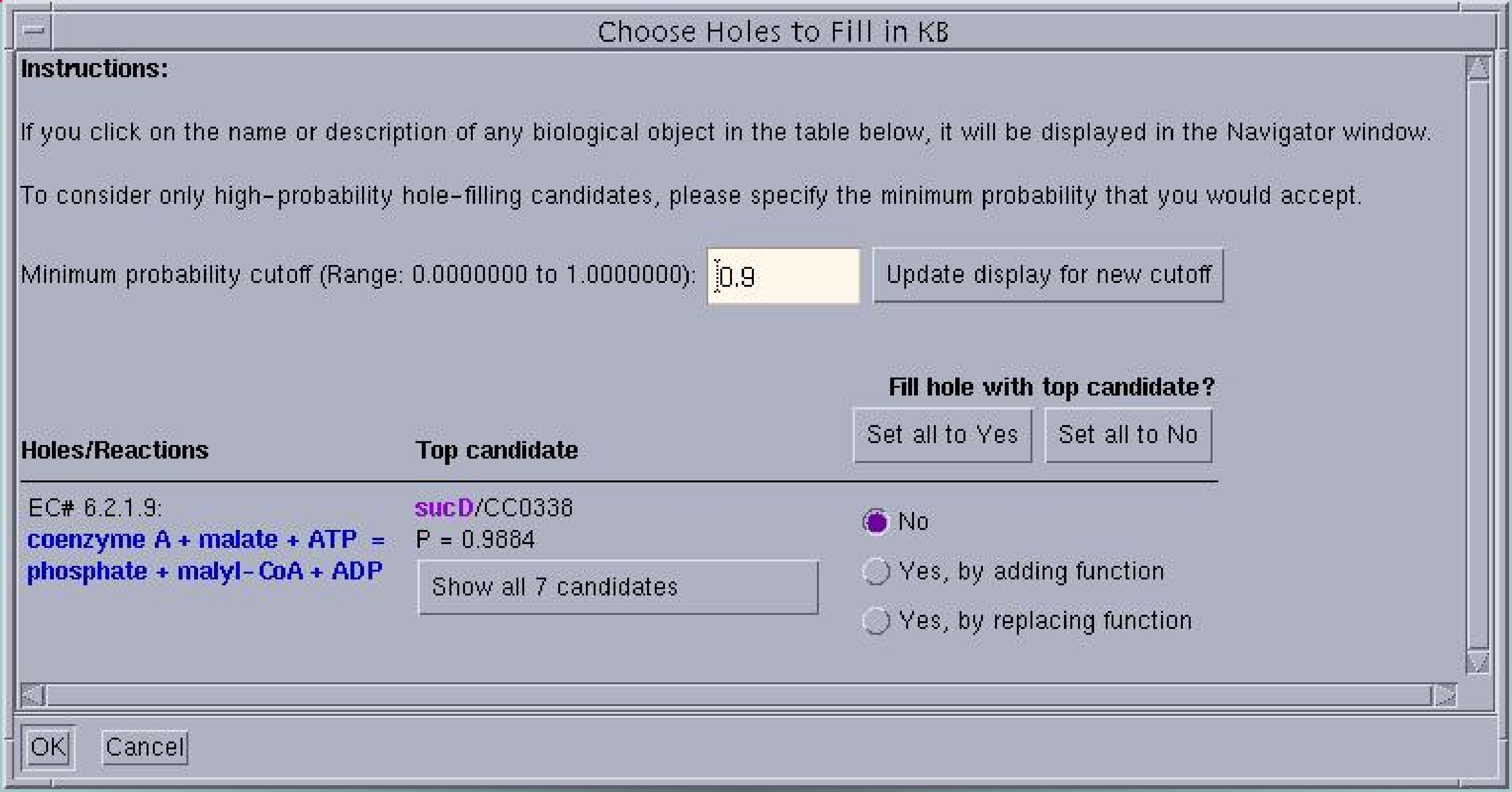
Wizard prompts user for training data source and for which holes to make predictions. Wizard runs Steps 1 & 2, then prompts user to complete Step 3.

Power-user mode

User must proceed through each step in order. Program still prompts user for required parameters, but each step must be completed before advancing to next step.



Step 3: Choose Holes to Fill in KB



Choose Holes to Fill in KB

Instructions:

If you click on the name or description of any biological object in the table below, it will be displayed in the Navigator window.

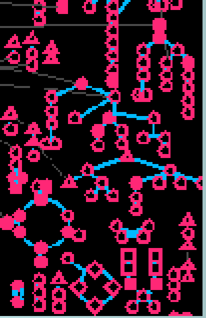
To consider only high-probability hole-filling candidates, please specify the minimum probability that you would accept.

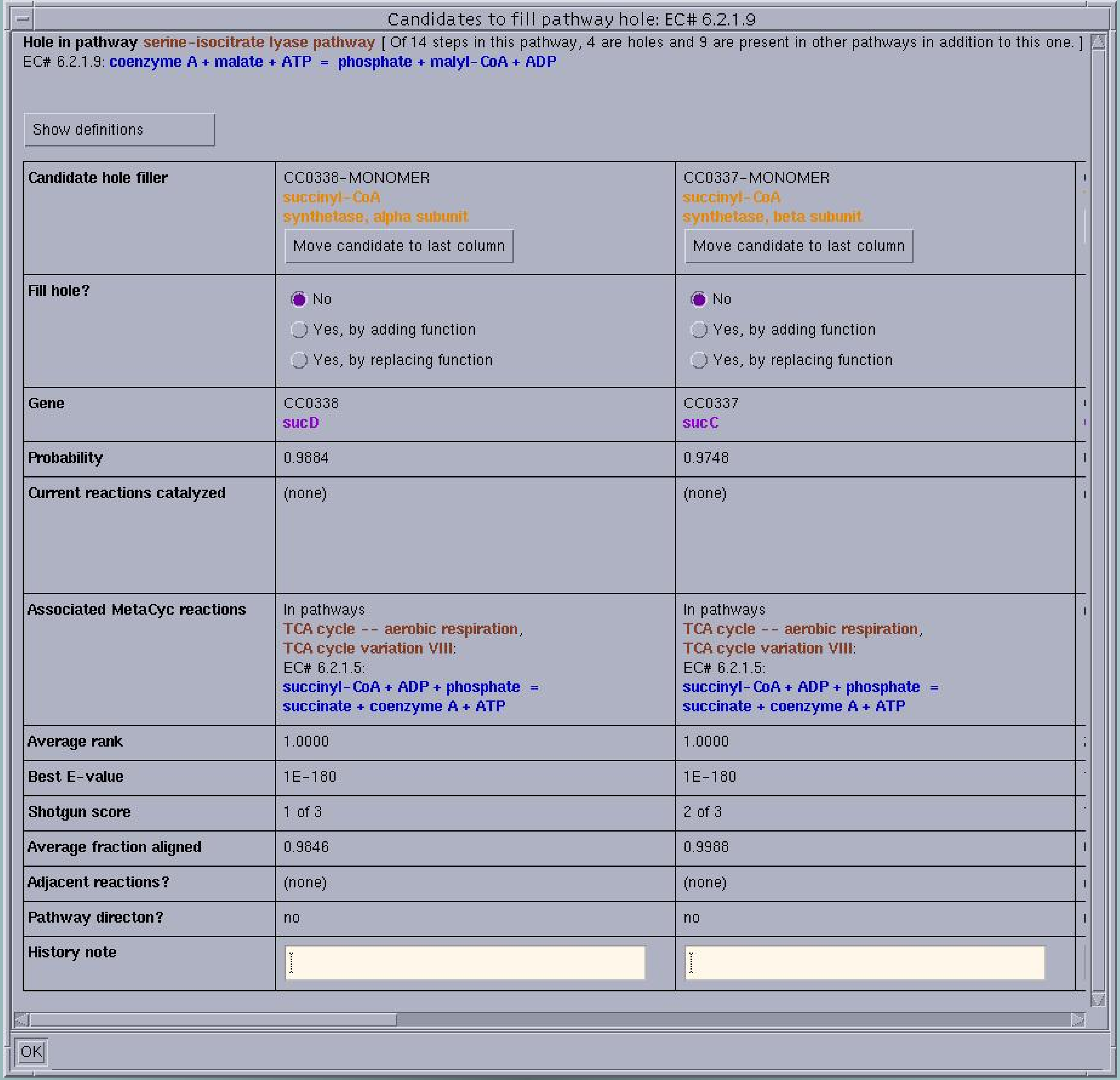
Minimum probability cutoff (Range: 0.0000000 to 1.0000000): 0.9 [Update display for new cutoff]

Fill hole with top candidate? [Set all to Yes] [Set all to No]

Holes/Reactions	**Top candidate**	
EC# 6.2.1.9: coenzyme A + malate + ATP = phosphate + malyl-CoA + ADP	sucD/CC0338 P = 0.9884 [Show all 7 candidates]	● No ○ Yes, by adding function ○ Yes, by replacing function

Biocyc.org



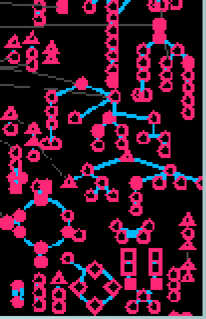


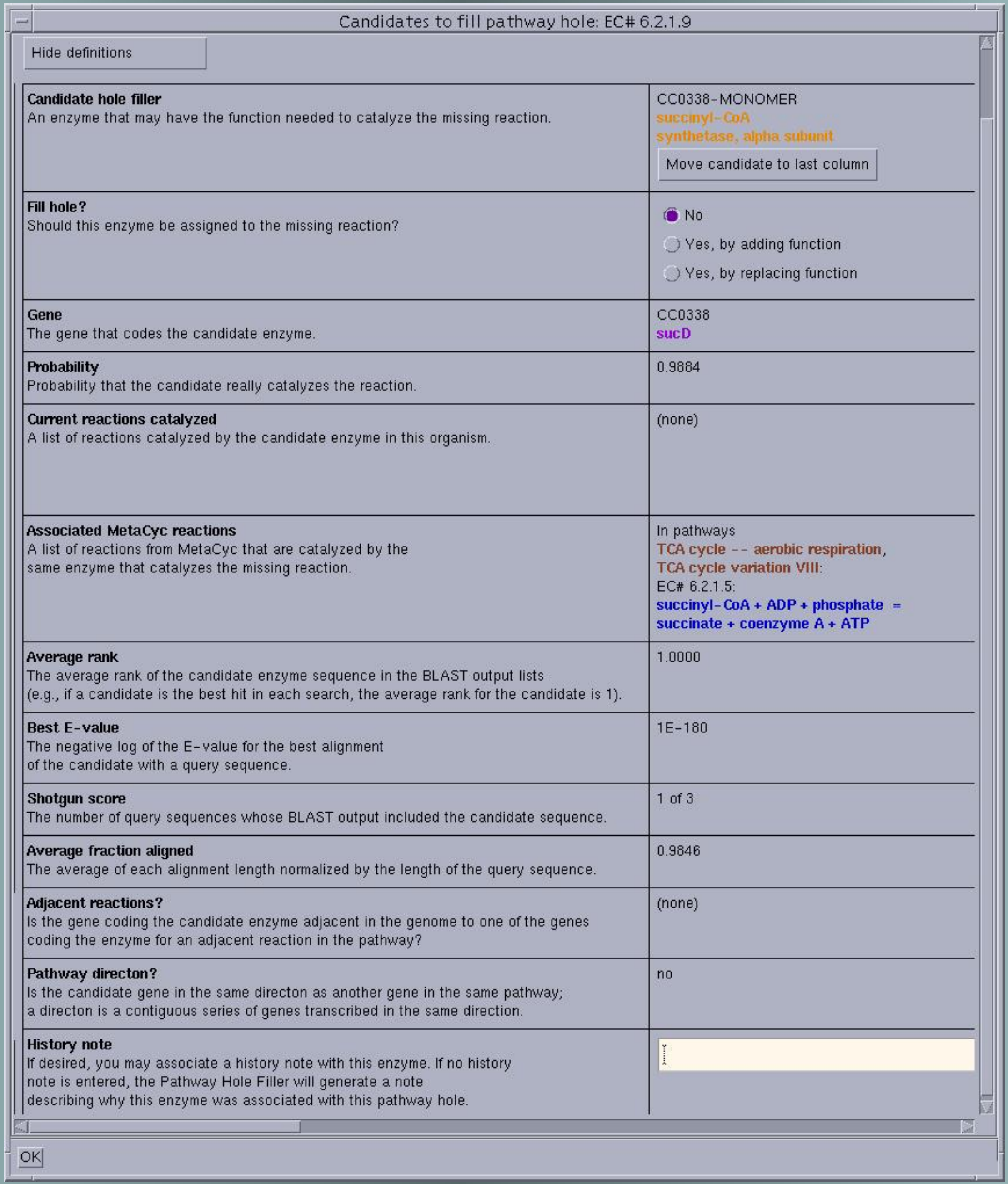
Candidates to fill pathway hole: EC# 6.2.1.9

Hole in pathway serine-isocitrate lyase pathway [Of 14 steps in this pathway, 4 are holes and 9 are present in other pathways in addition to this one.]
EC# 6.2.1.9: coenzyme A + malate + ATP = phosphate + malyl-CoA + ADP

Candidate hole filler	CC0338-MONOMER succinyl-CoA synthetase, alpha subunit Move candidate to last column	CC0337-MONOMER succinyl-CoA synthetase, beta subunit Move candidate to last column
Fill hole?	● No ○ Yes, by adding function ○ Yes, by replacing function	● No ○ Yes, by adding function ○ Yes, by replacing function
Gene	CC0338 sucD	CC0337 sucC
Probability	0.9884	0.9748
Current reactions catalyzed	(none)	(none)
Associated MetaCyc reactions	In pathways TCA cycle -- aerobic respiration, TCA cycle variation VIII: EC# 6.2.1.5: succinyl-CoA + ADP + phosphate = succinate + coenzyme A + ATP	In pathways TCA cycle -- aerobic respiration, TCA cycle variation VIII: EC# 6.2.1.5: succinyl-CoA + ADP + phosphate = succinate + coenzyme A + ATP
Average rank	1.0000	1.0000
Best E-value	1E-180	1E-180
Shotgun score	1 of 3	2 of 3
Average fraction aligned	0.9846	0.9988
Adjacent reactions?	(none)	(none)
Pathway directon?	no	no
History note		

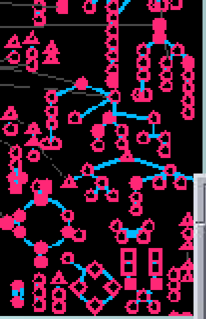
Biocyc.org

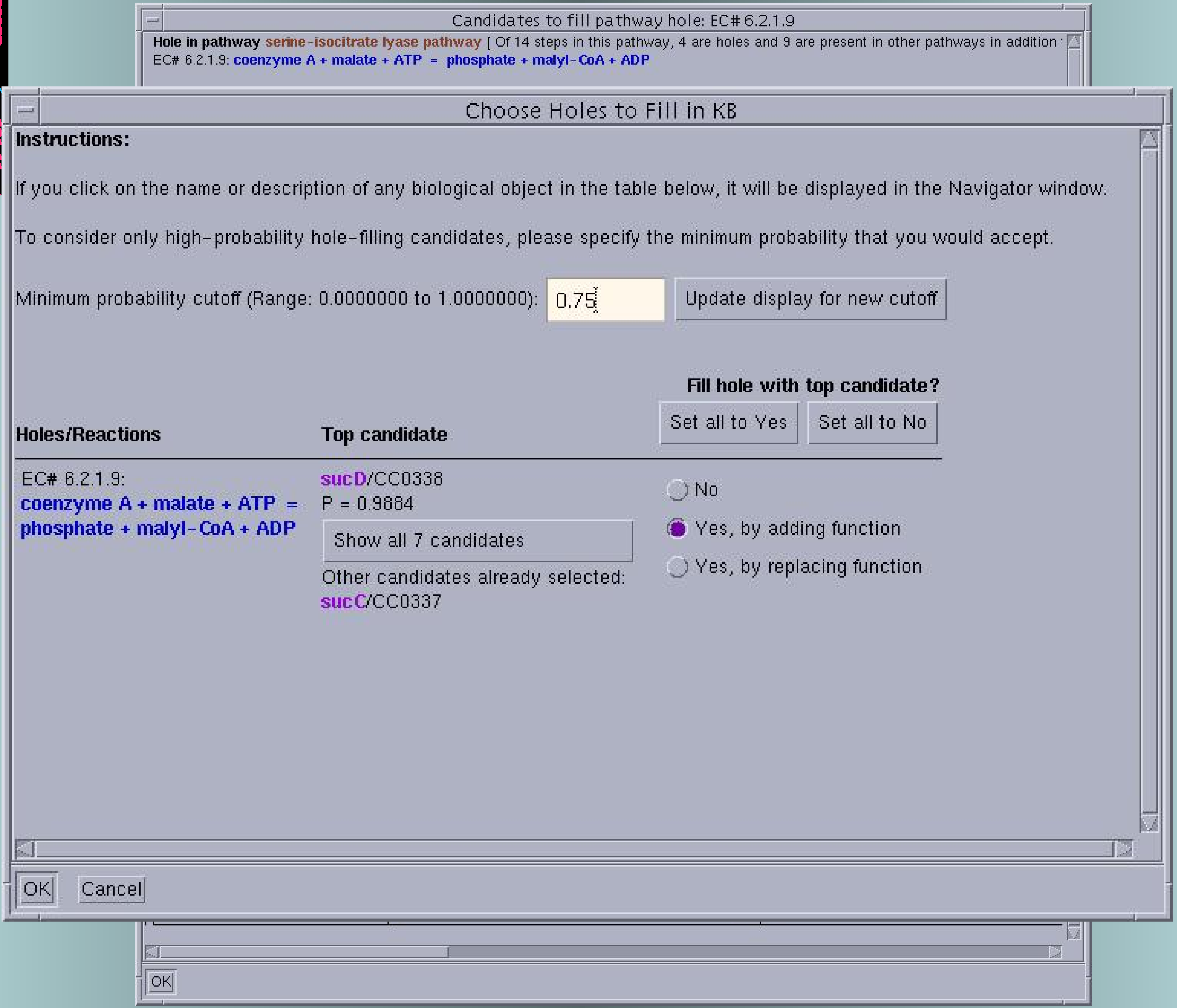




Candidates to fill pathway hole: EC# 6.2.1.9

Candidate hole filler An enzyme that may have the function needed to catalyze the missing reaction.	CC0338-MONOMER succinyl-CoA synthetase, alpha subunit Move candidate to last column
Fill hole? Should this enzyme be assigned to the missing reaction?	● No ○ Yes, by adding function ○ Yes, by replacing function
Gene The gene that codes the candidate enzyme.	CC0338 sucD
Probability Probability that the candidate really catalyzes the reaction.	0.9884
Current reactions catalyzed A list of reactions catalyzed by the candidate enzyme in this organism.	(none)
Associated MetaCyc reactions A list of reactions from MetaCyc that are catalyzed by the same enzyme that catalyzes the missing reaction.	In pathways TCA cycle -- aerobic respiration, TCA cycle variation VIII: EC# 6.2.1.5: succinyl-CoA + ADP + phosphate = succinate + coenzyme A + ATP
Average rank The average rank of the candidate enzyme sequence in the BLAST output lists (e.g., if a candidate is the best hit in each search, the average rank for the candidate is 1).	1.0000
Best E-value The negative log of the E-value for the best alignment of the candidate with a query sequence.	1E-180
Shotgun score The number of query sequences whose BLAST output included the candidate sequence.	1 of 3
Average fraction aligned The average of each alignment length normalized by the length of the query sequence.	0.9846
Adjacent reactions? Is the gene coding the candidate enzyme adjacent in the genome to one of the genes coding the enzyme for an adjacent reaction in the pathway?	(none)
Pathway directon? Is the candidate gene in the same directon as another gene in the same pathway; a directon is a contiguous series of genes transcribed in the same direction.	no
History note If desired, you may associate a history note with this enzyme. If no history note is entered, the Pathway Hole Filler will generate a note describing why this enzyme was associated with this pathway hole.	





Candidates to fill pathway hole: EC# 6.2.1.9

Hole in pathway serine-isocitrate lyase pathway [Of 14 steps in this pathway, 4 are holes and 9 are present in other pathways in addition

EC# 6.2.1.9: coenzyme A + malate + ATP = phosphate + malyl-CoA + ADP

Choose Holes to Fill in KB

Instructions:

If you click on the name or description of any biological object in the table below, it will be displayed in the Navigator window.

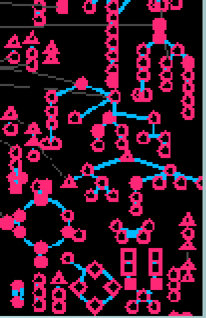
To consider only high-probability hole-filling candidates, please specify the minimum probability that you would accept.

Minimum probability cutoff (Range: 0.0000000 to 1.0000000): 0.75

Fill hole with top candidate?

Holes/Reactions	**Top candidate**	
EC# 6.2.1.9: coenzyme A + malate + ATP = phosphate + malyl-CoA + ADP	sucD/CC0338 P = 0.9884 Show all 7 candidates Other candidates already selected: sucC/CC0337	○ No ● Yes, by adding function ○ Yes, by replacing function

Biocyc.org



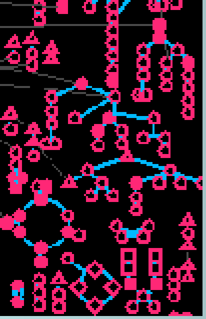
Output from Pathway Hole Filler - from “Prepare Training Data” step

ROOT/ptools-local/pgdbs/user/ORGIDcyc/VERSION/data/

(e.g., ROOT/ptools-local/pgdbs/user/caulocyc/1.0/data/)

- rxn-list = data retrieved from ORGID for calculating training data
- priors/ = directory containing training data that is loaded when using existing data from ORGID
- These files contain the training data computed in Step 1. If either file is available, the user may use “existing” training data in Step 1.

* Each file is overwritten each time you run this step.



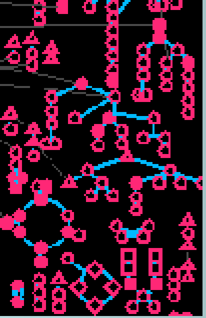
Output from Pathway Hole Filler
- from “Identify and Evaluate Candidates” step

ROOT/ptools-local/pgdbs/user/ORGIDcyc/VERSION/reports/

(e.g., ROOT/ptools-local/pgdbs/user/caulocyc/1.0/reports/)

- ***ORGID_filled-holes.html*** = the list of holes that user selected to fill in the KB in Step 3.
- *ORGIDholesX-Y.html* **(e.g., CAULOholes0-10.html)**
- blasterrors.log = log of each rxn describing whether or not any candidates were found
- hole-data = file containing data found for each rxn, used to generate list in “Choose holes to fill in KB” dialogue. If this file is available, step 3 can be initiated without repeating Step 2.

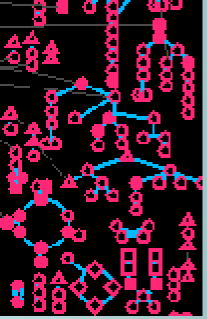
* Each file is overwritten each time you run this step.



Reference for the Pathway Hole Filler

Green, ML and Karp, PD.

A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.
BMC Bioinformatics 2004, 5:76.



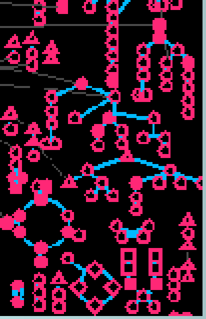
Pathway Hole Filler Demo (1)

Prerequisites:

- HpyCyc installed
- BLAST installed and working
- For EcoCyc, the data/priors/ directory needed

Demo:

- Using Power User mode, to save time
- Select HpyCyc
- Refine->PHF->Step 1: Prepare Training Data
- In popup, select HpyCyc and 2-3 reactions



Pathway Hole Filler Demo (2)

- once more:

  Refine->PHF->Step 1: Prepare Training Data
- In popup, select EcoCyc and say Yes to

  use existing Training Data
- Refine->PHF->Step 2: Identify Candidates
  - In popup, select Pathways from a List
  - Select Pyridnucsyn-Pwy
- Refine->PHF->Step 3: Choose Holes to Fill in KB