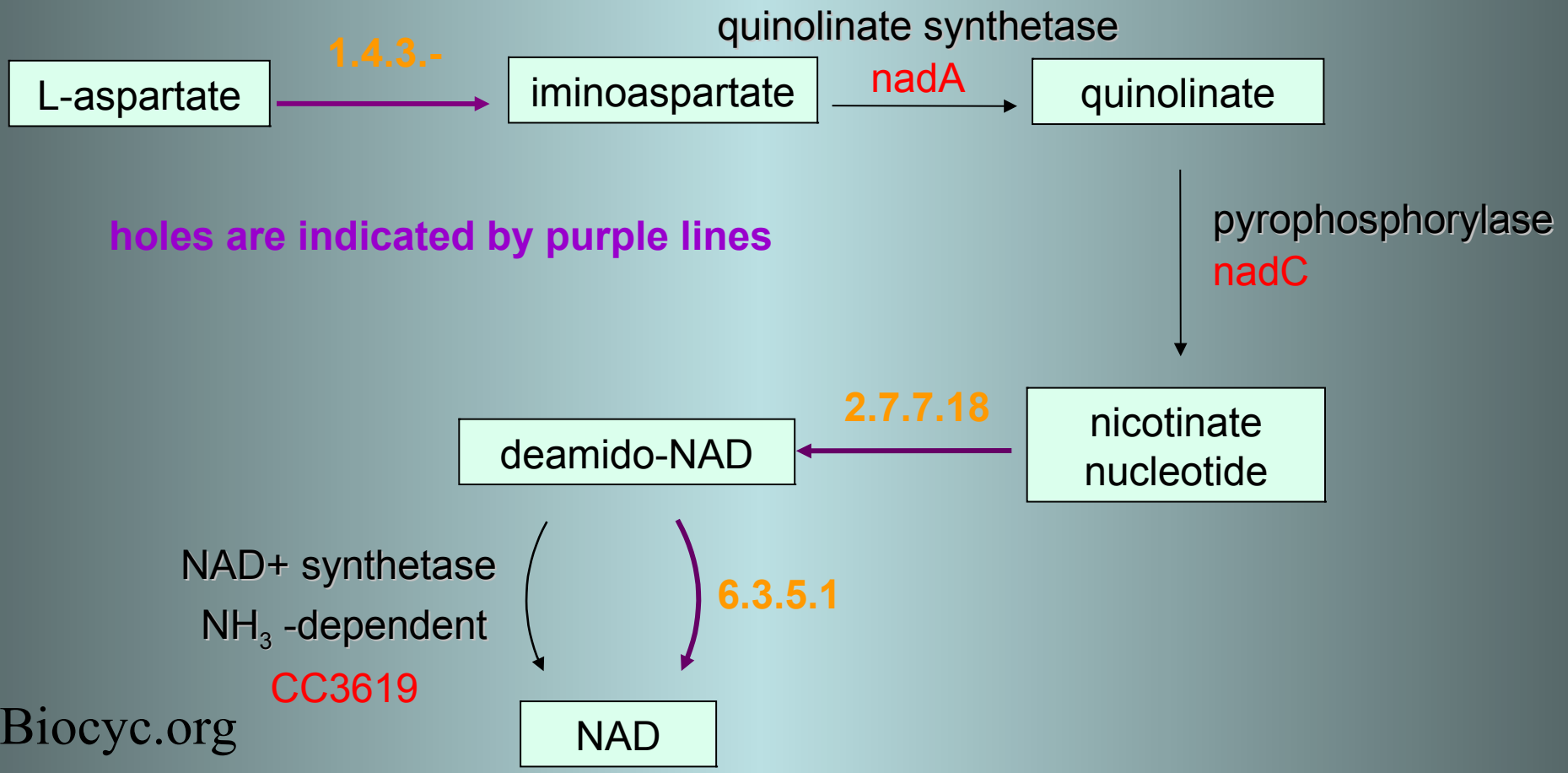
# Identify Pathway Hole Fillers

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

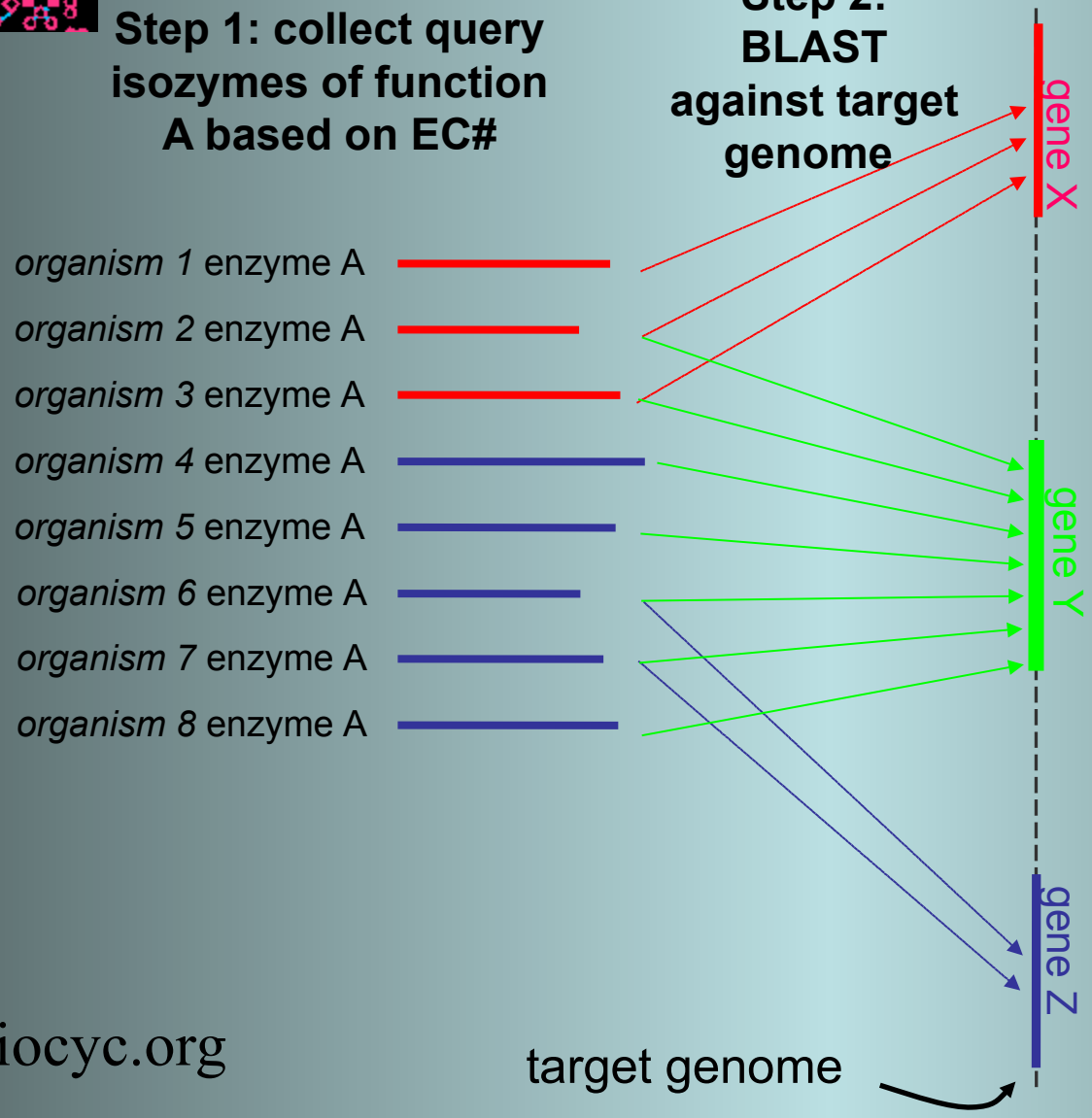



Biocyc.org

# 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**

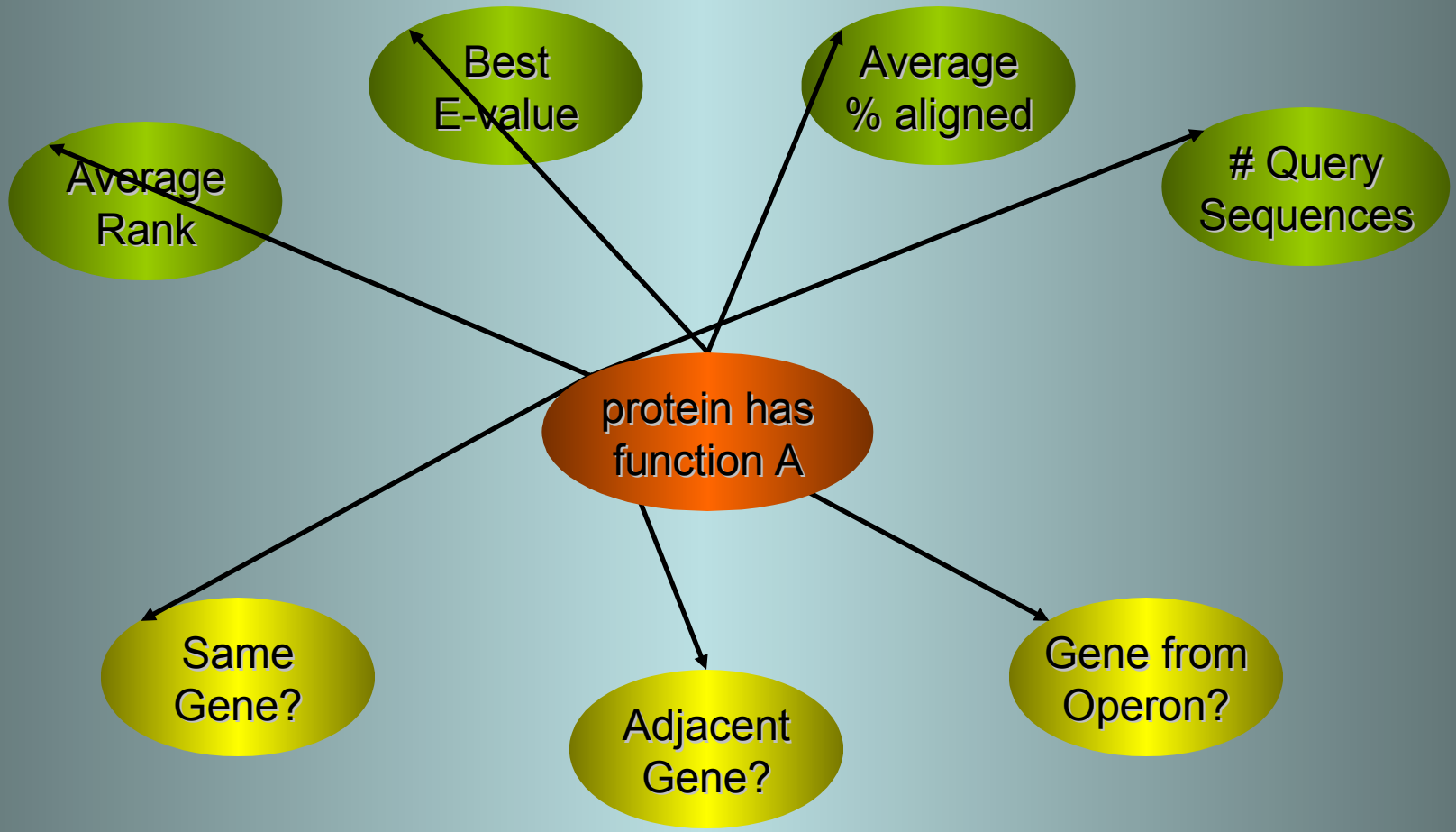
**Gene Z**

Biocyc.org

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

- Best E-value
- Avg. rank of candidate sequence in BLAST output
- Avg. length % 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?

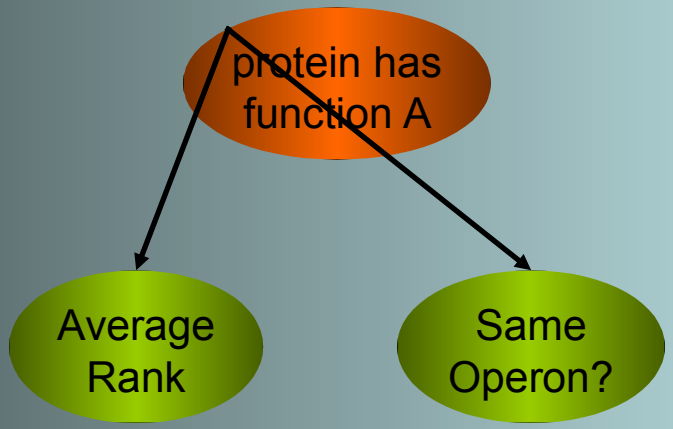
# Use Bayesian classifier to evaluate candidates



# Computing P(has function)

Apply Bayes' rule:

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




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

e.g., Same operon?

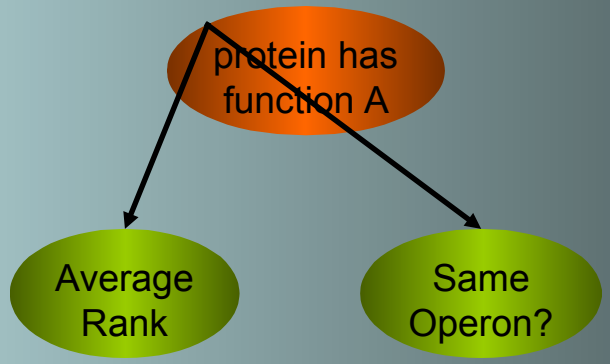
| In operon? | True hit Has-Fn(A) | False hit ~Has-Fn(A) |
|---|---|---|
| yes | 0.24 (TP) | 0.04 (FP) |
| no | 0.76 (FN) | 0.96 (TN) |

# 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)

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

$$0.77$$

Biocyc.org

# Steps that must be completed before running the Pathway Hole Filler

- Install BLAST executable (see Installation instructions)
- Prepare BLAST protein db for the PGDB
  - Need FASTA format genome nucleotide sequence. (If only ESTs are available, see User Guide regarding Prepare BLAST Reference Data->Protein from ESTs)
- In general, the more pathways in your PGDB, the more candidates the pathway hole filler will have to find

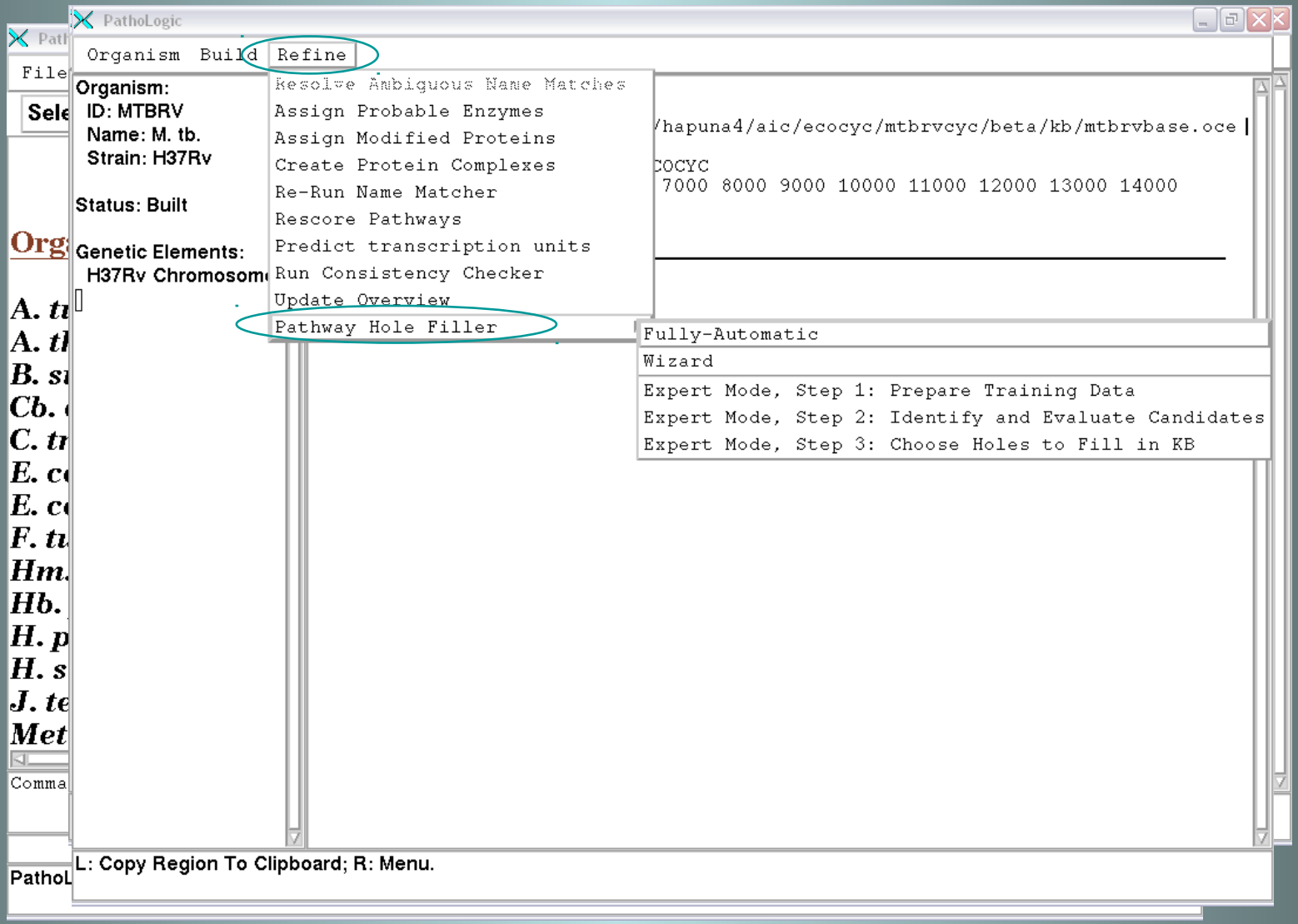
## *Conceptual stages of the pathway hole filler*

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



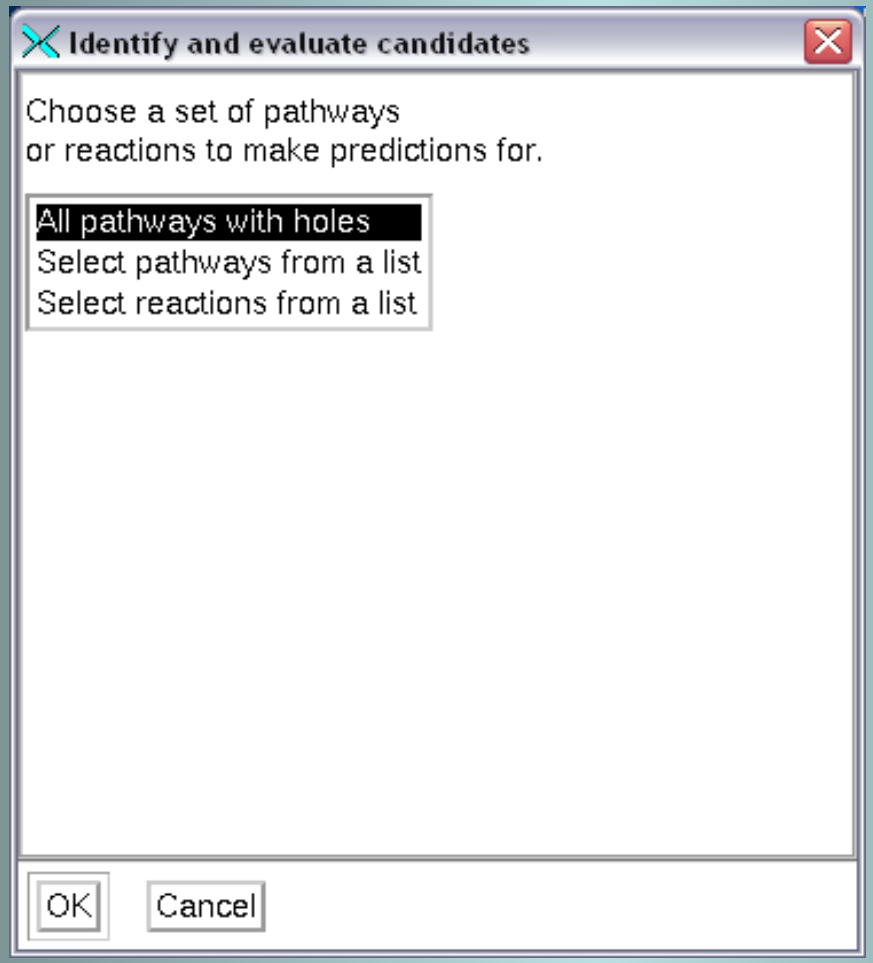

Biocyc.org

# Step 1: Prepare Training Data

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

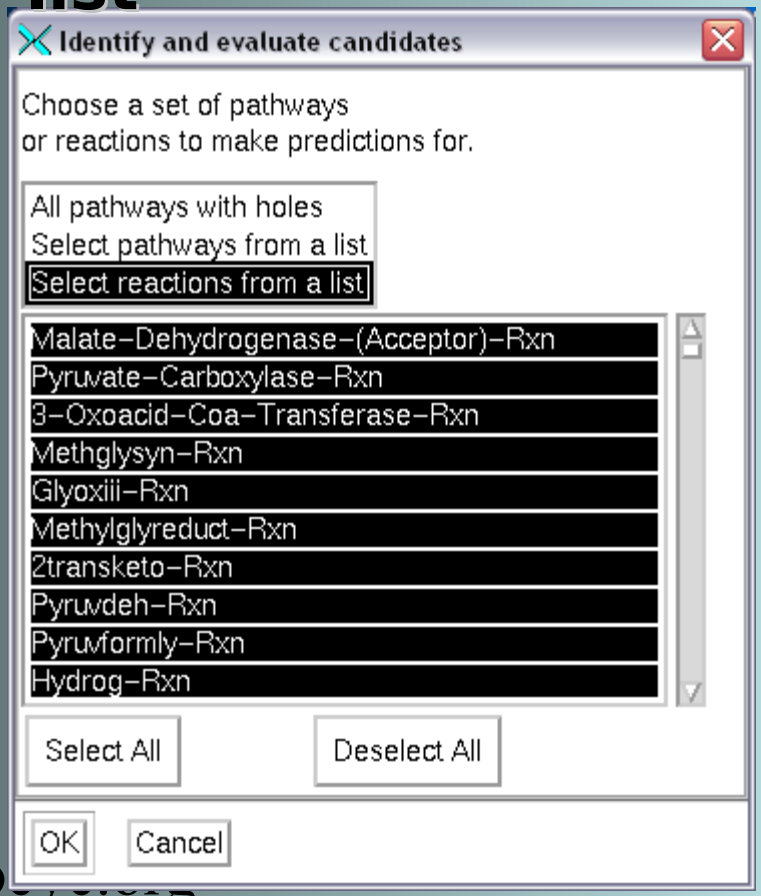
- Once Step 1 has been completed, the training data are saved and can be reused (even in another Pathway Tools session).
- If using existing data from *E. coli* the training data are based on data from the literature.

# Step 2: Identify & Evaluate Candidates…

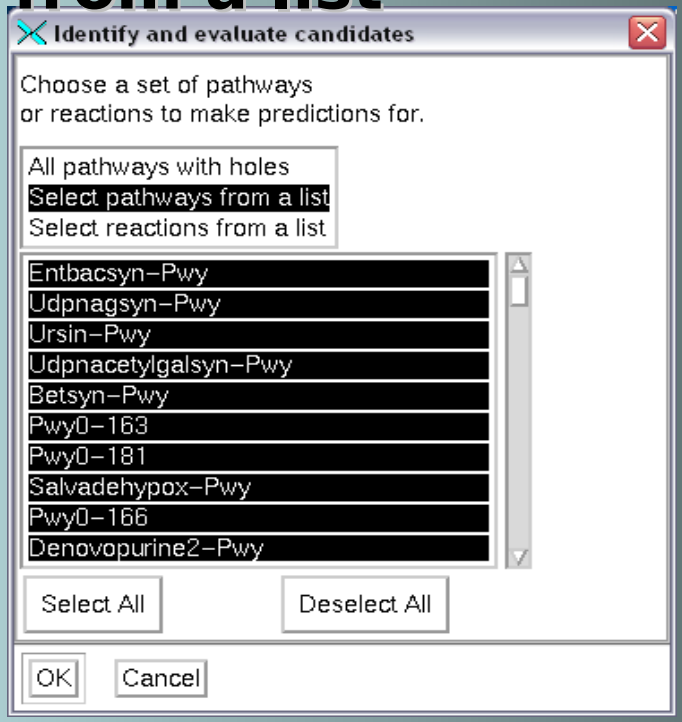


# Step 2: Identify & Evaluate Candidates

## Select reactions from a list




## Select pathways from a list



# 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

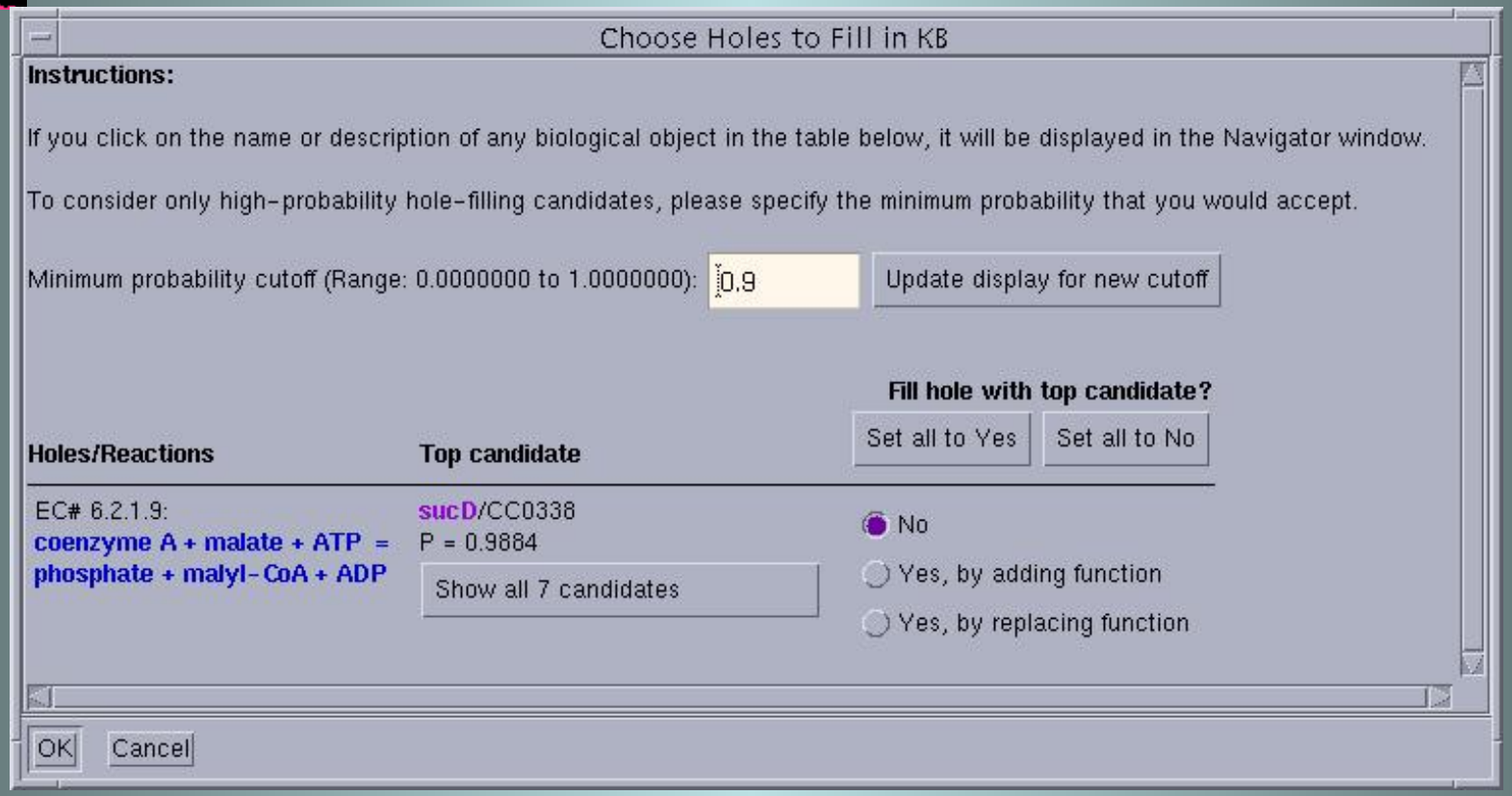
*Wizard*

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




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**

Show definitions

| | | |
|---|---|---|
| **Candidate hole filler** | CC0338-MONOMER<br>succinyl-CoA<br>synthetase, alpha subunit<br>Move candidate to last column | CC0337-MONOMER<br>succinyl-CoA<br>synthetase, beta subunit<br>Move candidate to last column |
| **Fill hole?** | (•) No<br>( ) Yes, by adding function<br>( ) Yes, by replacing function | (•) No<br>( ) Yes, by adding function<br>( ) Yes, by replacing function |
| **Gene** | CC0338<br>sucD | CC0337<br>sucC |
| **Probability** | 0.9884 | 0.9748 |
| **Current reactions catalyzed** | (none) | (none) |
| **Associated MetaCyc reactions** | In pathways<br>TCA cycle -- aerobic respiration,<br>TCA cycle variation VIII:<br>EC# 6.2.1.5:<br>succinyl-CoA + ADP + phosphate =<br>succinate + coenzyme A + ATP | In pathways<br>TCA cycle -- aerobic respiration,<br>TCA cycle variation VIII:<br>EC# 6.2.1.5:<br>succinyl-CoA + ADP + phosphate =<br>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** | | |

OK

Biocyc.or

Candidates to fill pathway hole: EC# 6.2.1.9

Hide definitions

| | |
|---|---|
| **Candidate hole filler**<br>An enzyme that may have the function needed to catalyze the missing reaction. | CC0338-MONOMER<br>succinyl-CoA synthetase, alpha subunit<br>Move candidate to last column |
| **Fill hole?**<br>Should this enzyme be assigned to the missing reaction? | (•) No<br>( ) Yes, by adding function<br>( ) Yes, by replacing function |
| **Gene**<br>The gene that codes the candidate enzyme. | CC0338<br>sucD |
| **Probability**<br>Probability that the candidate really catalyzes the reaction. | 0.9884 |
| **Current reactions catalyzed**<br>A list of reactions catalyzed by the candidate enzyme in this organism. | (none) |
| **Associated MetaCyc reactions**<br>A list of reactions from MetaCyc that are catalyzed by the same enzyme that catalyzes the missing reaction. | In pathways<br>TCA cycle -- aerobic respiration, TCA cycle variation VIII:<br>EC# 6.2.1.5:<br>succinyl-CoA + ADP + phosphate = succinate + coenzyme A + ATP |
| **Average rank**<br>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**<br>The negative log of the E-value for the best alignment of the candidate with a query sequence. | 1E-180 |
| **Shotgun score**<br>The number of query sequences whose BLAST output included the candidate sequence. | 1 of 3 |
| **Average fraction aligned**<br>The average of each alignment length normalized by the length of the query sequence. | 0.9846 |
| **Adjacent reactions?**<br>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?**<br>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**<br>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. | |

OK

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

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 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<br>P = 0.9884<br>Show all 7 candidates<br>Other candidates already selected: sucC/CC0337 | ○ No<br>● Yes, by adding function<br>○ Yes, by replacing function |

OK Cancel

OK

Biocyc.org

# 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 (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

# 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