PathoLogic: More about Matching Enzymes to Reactions

Inputs

- MetaCyc is the primary reference PGDB. (Most) name/reaction associations in MetaCyc will be available to the name matcher
- You can specify additional reference PGDBs using the *Organism -> Specify Reference PGDB(s)* menu item – useful if there is a manually curated PGDB for a closely related organism.
- Several additional name/reaction mapping files.

Mapping Files

- Allow you to specify additional name-reaction mappings not present in PGDBs
- aic-export/pathway-tools/pathologic/VERSION/data/enzyme-mappings.dat: global mappings provided by us; updated with new PTools release
- ptools-local/local-enzyme-mappings.dat: your local mappings; apply to all new PGDBs; persist between PTools upgrades
- ptools-local/pgdbs/user/ORGIDcyc/VERSION/input/enzyme-mappings.dat: for current PGDB only

Mapping File Format

- Tab-delimited; two required columns, one optional
- Column 1: name
- Column 2: space-separated list of reactions associated with name
- Column 3 (optional): flag indicating whether name is ambiguous (T/NIL)

Overview of name matching

- **The name matcher runs in three phases:**
  - Phase I: Build a table indexing the names in MetaCyc (and other ref. PGDBs) and the associated reactions; names checked for ambiguity
  - Phase II: Look up protein names from the annotated genome in the table
  - Phase III: Analyze nonmatching enzymes – look for “probable enzymes” and possible matches

Phase I: Build Index

- **Protein function names can be stored in multiple places in MetaCyc. The name matcher indexes names from:**
  - reaction frames (e.g., official names assigned by EC)
  - enzymatic reaction frames
  - enzyme frames (provided that the enzyme is monofunctional and the name contains “ase”
  - mapping files (described above)
- **Names in gene frames are not indexed.**

Phase I: Build Index

- **Each name can be associated with multiple reactions.**
- **In some cases, the name is ambiguous. A pair of reactions having the same name are ambiguous if:**
  - the reactions' EC numbers (if any) don't match (partial match is okay, e.g., 1.2.3.- with 1.2.3.4)
  - no enzyme in MetaCyc catalyzes both reactions
- **Ambiguous matches are presented to the user for review. (“Assign Probable Enzymes”)**

Phase II: Look up names

- **In the simplest case, a protein has one function with one name. If the name exactly matches a name in the table, associate the protein with the reactions for that name. (Spaces and punctuation are ignored.)**
- **Some proteins have multiple functions with multiple names. How are they handled?**
  - Multiple functions are treated separately – each can give a matching set of reactions.
  - Multiple names for a function are considered together – the first name that matches determines the reaction set.

But wait, there's more!

- **The name matcher doesn't just check the exact name given in the annotation file. If the original name doesn't match, it tries a variety of “alternative” names:**
  - remove common prefixes and suffixes, such as “putative”, “probable”, “hypothetical”, “homolog”, “family protein”, etc.
  - remove “subunit ___”, “small chain”, etc. (But see the Create Protein Complexes task)
  - remove some “gene name”-like names: e.g., “xyzA”, short parenthesized words

Phase III: Nonmatching names

- **If a name can't be matched, even in an alternative form, we try to decide whether it is a “likely metabolic enzyme”. This is true if:**
  - the name contains “ase”
  - the name doesn't contain “RNA”, except “tRNA”
  - the name doesn't match a list of nonmetabolic enzyme names (aic-export/pathway-tools/pathologic/VERSION/data/metabolic-enzyme-ruleout-words.dat)
  - the name doesn't match a list of nonspecific enzyme names (aic-export/pathway-tools/pathologic/VERSION/data/nonspecific-enzyme-names.dat)

Phase III: Nonmatching names

- Likely metabolic enzymes can be reviewed in the "Assign Probable Enzymes" task under "Refine"
- Right click on an enzyme to get information about that enzyme, include a list of suggested reactions.
- Suggested reactions are found by approximate matching to MetaCyc names
- You can also split an enzyme name into separate names, flag for future research, or reject (nonmetabolic / nonspecific)
- See also name matching report: ptools-local/pgdbs/user/ORGIDcyc/VERSION/reports/name-matching-report.txt

How it fits together

- PathoLogic can use function names, EC numbers, and GO molecular function annotations to match enzymes to reactions
- A single protein may have any or all of these annotation types

How it fits together

- EC number annotations for a protein are considered first. A single EC number can match one or more MetaCyc reactions.
- If no EC numbers are present, or if they don't match anything in MetaCyc (e.g., partial ECs), name matches are considered.
- GO annotations are considered last. Additional matches can be added, even if matches were found using ECs or names.