Large-Scale Metabolic Network Alignment: MetaCyc and KEGG

Tomer Altman

Bioinformatics Research Group

SRI International

taltman@ai.sri.com

Problem Motivation

- There are an increasing number of ‘encyclopedic’ metabolic networks, or reaction databases
- KEGG and MetaCyc, plus Rhea, BRENDA, and GO
- A natural question to ask is, “what is similar / different between them?”
- There has been some linking of MetaCyc compounds to KEGG, but none for reactions up until 2009

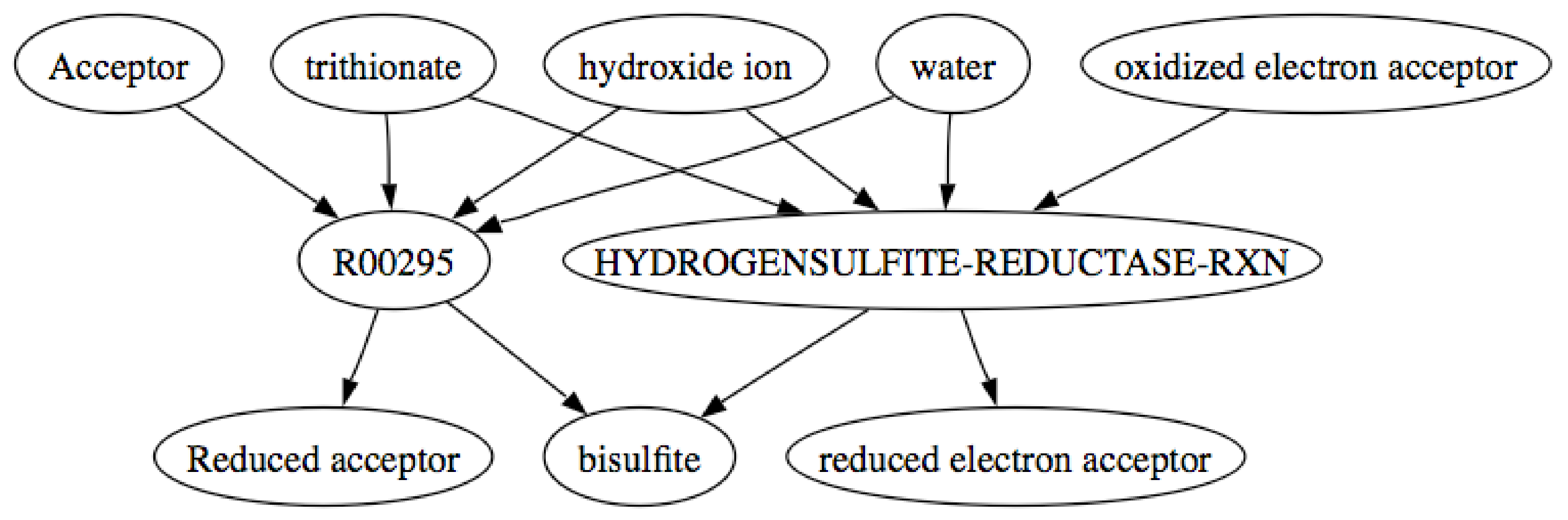
Challenges with Mapping Objects

- Multiple aspects to compare (name, chemical structure, reaction substrates, external identifiers)
- Inexact naming
- Inexact structures (different specificity of stereocenters)
- Inexact description of reactions (classes vs. instances, proton-balancing)
- How to combine the evidence in a logical fashion

Compound Evidence

- Curated MetaCyc links to KEGG
- Name matching
- PubChem identifier mapping (used for ChEBI as well)
- Molecular Fingerprint Tanimoto Similarity Coefficient
- InChI string comparison
- Exact Sub-Structure Match (no stereochemistry)
- ‘All-but-one’ inference

Compound Prediction Detail: ‘All-but-one’

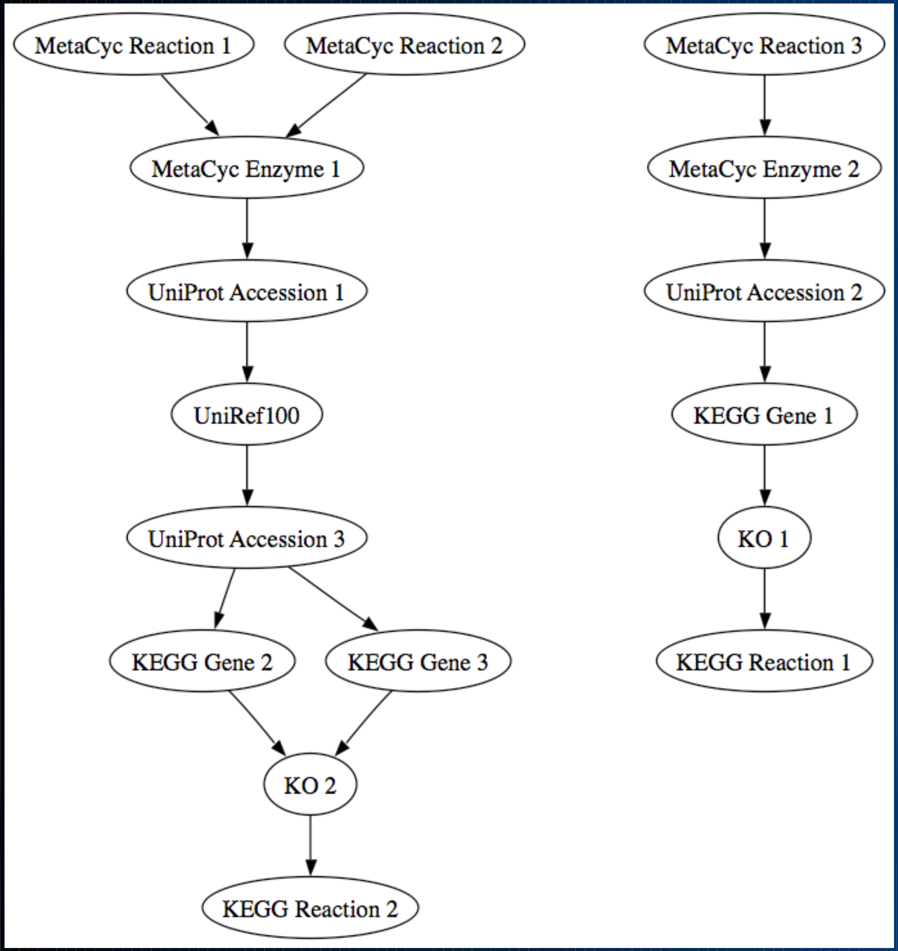

Acceptor
trithionate
hydroxide ion
water
oxidized electron acceptor
R00295
HYDROGENSULFITE-REDUCTASE-RXN
Reduced acceptor
bisulfite
reduced electron acceptor

- **Most of the compounds between these two reactions are the same**
- **Class vs. instance, and naming issues lead to unknown match between “acceptor” and “oxidized electron acceptor”**

Reaction Evidence

- EC Numbers
- UniProt Accession Numbers
- Name matches (gleaned from associated objects)
- Exact equation match
- Inexact equation match (cosine similarity)

Reaction Prediction Detail: UniProt Mapping



- Use UniProt Accession numbers to map the enzymes in MetaCyc and KEGG to one another
- Use UniRef 90 or 100 to map “the same protein” when not exact same Accession Number

From Evidence to Prediction

- **First approach involved bootstrapping the mapping by means of an ad-hoc algorithm that was tuned to be very conservative, and subsequent validation by curation staff**
- **Currently a machine learning approach to evaluating all of the features shared between reactions in Kegg and MetaCyc is being developed with collaborators at Stanford**
  - Evaluate features for information content
  - Implement as Naïve Bayes, Logistic Regression, SVM, etc. to determine method with greatest predictive power
  - Classify unmapped data with hierarchical clustering (i.e., unsupervised learning)
  - Provide as general algorithm for comparing reaction databases

Current Status and Future Work

- ### MetaCyc reactions with links to KEGG (~##%)
- ### MetaCyc compounds with links to KEGG (>## %)
- Analyzing unmatched content of KEGG and MetaCyc for algorithm improvement and focused curation
- Development of new features for machine learning analysis

Acknowledgements

- Peter Karp
- Douglas Brutlag
- Anamika Kothari
- Carol Fulcher
- Ron Caspi
- Dan Davison
- Luciana Ferrer
- Joseph Dale

MetaCyc.org