Resolving ontology mappings using Boomer

Jim Balhoff & Chris Mungall

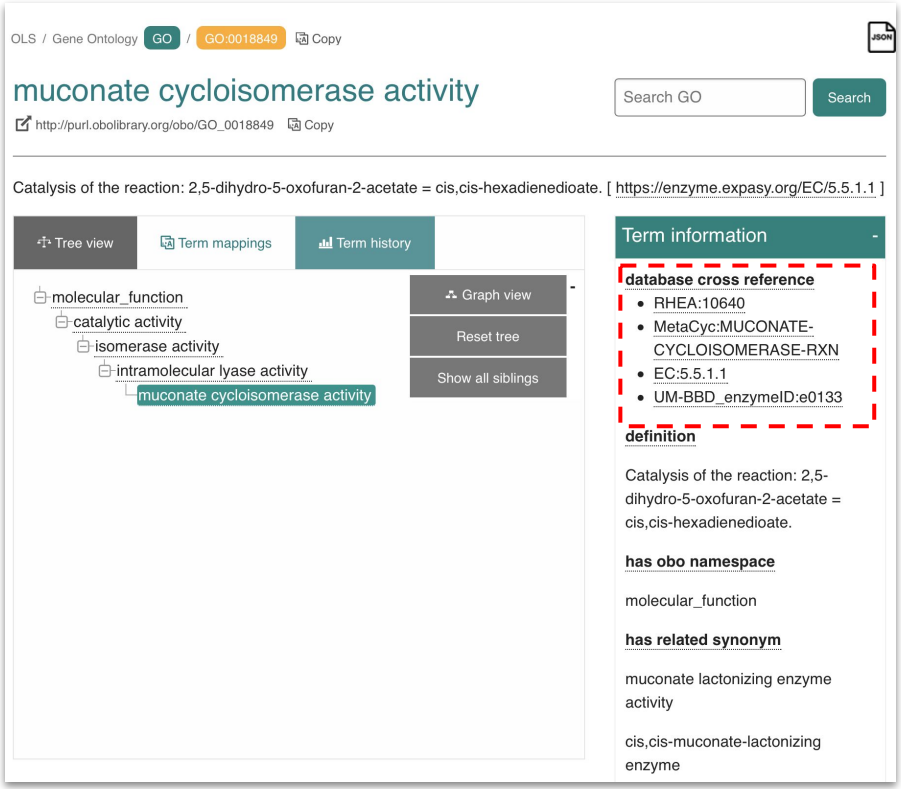
WSBO-2021: Workshop on Synergizing Biomedical Ontologies
July 15, 2021

renci

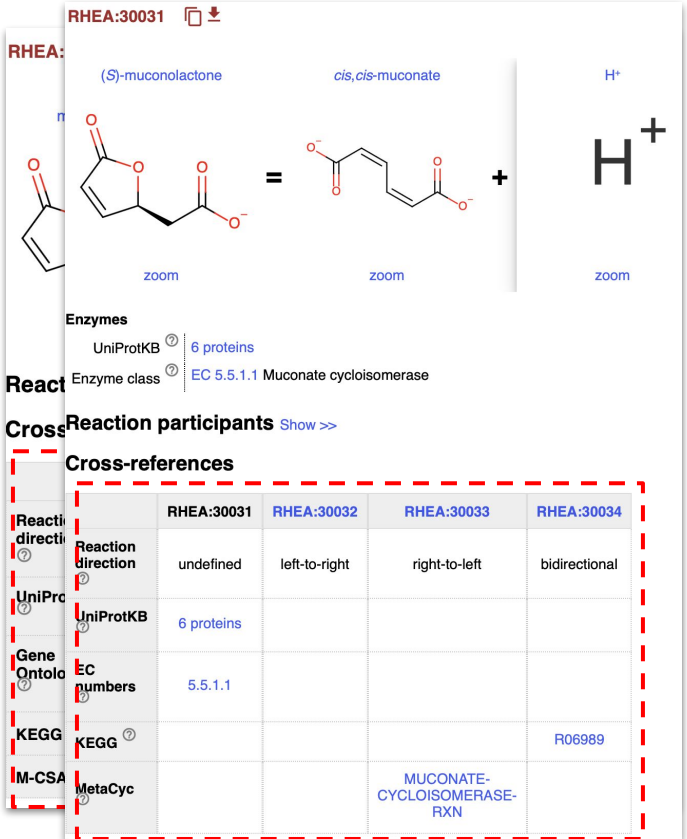
BERKELEY LAB

Mutual cross-references between terminologies

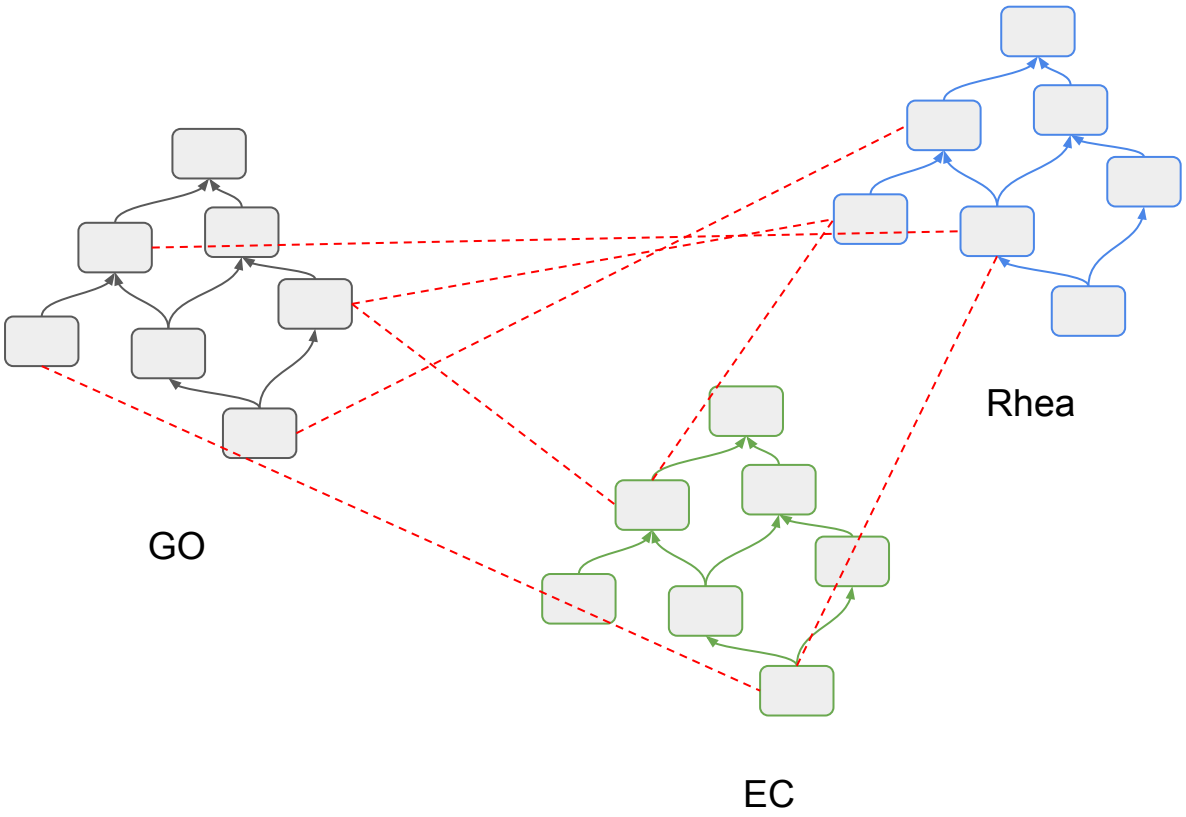
Gene Ontology



Rhea reaction DB



Mapping across hierarchies

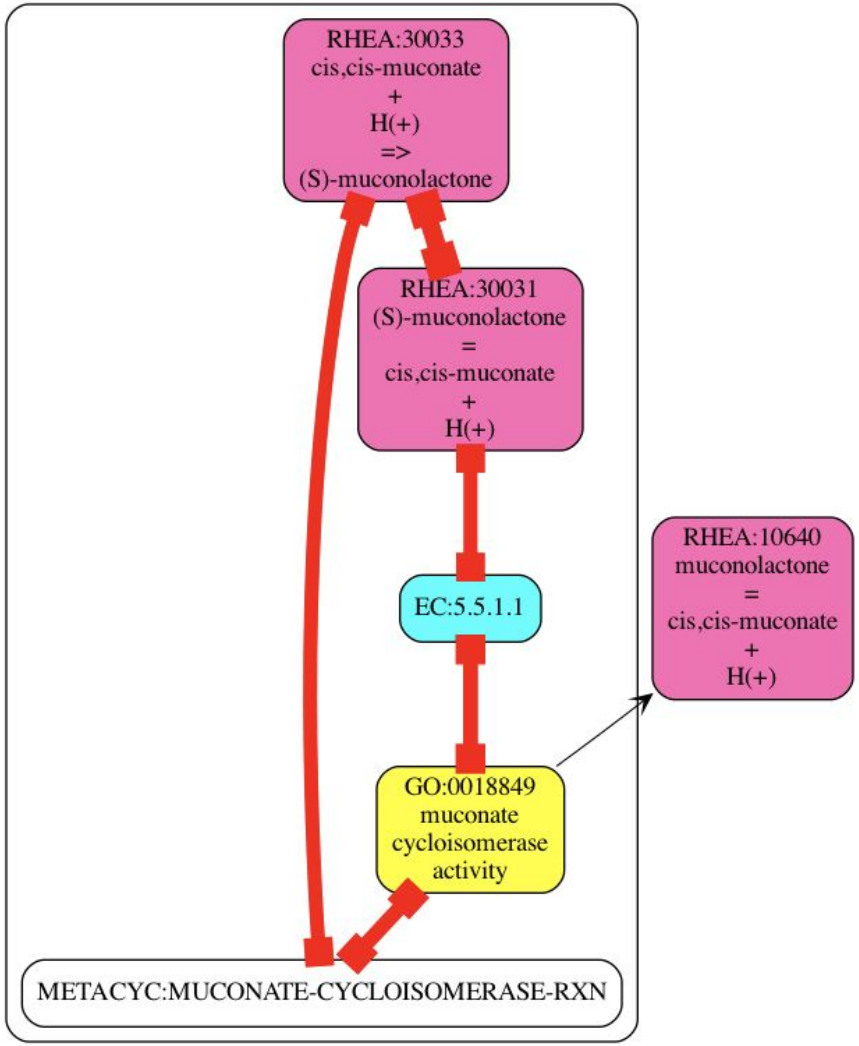



```
RHEA:10104    EC:1.14.11.15
RHEA:10108    EC:1.1.1.248
RHEA:10112    EC:4.2.1.22
RHEA:10116    EC:1.1.1.44
RHEA:10116    EC:1.1.1.351
RHEA:17176    METACYC:GSPAMID-RXN
RHEA:17236    METACYC:FUCISOM-RXN
RHEA:26155    METACYC:LUMAZINESYN-RXN
RHEA:10041    REACTOME:R-HSA-176606.2
RHEA:10041    REACTOME:R-HSA-8953499.2
RHEA:10117    REACTOME:R-HSA-71299.3
GO:0000009    EC:2.4.1.232
GO:0000010    RHEA:20836
GO:0000016    EC:3.2.1.108
GO:0000016    METACYC:LACTASE-RXN
GO:0000016    RHEA:10076
GO:0000025    METACYC:MALTOSECAT-PWY
GO:0000034    EC:3.5.4.2
GO:0000906    METACYC:LUMAZINESYN-RXN
GO:0000034    RHEA:23688
GO:0000048    EC:2.3.2.12
GO:0000105    METACYC:HISTSYN-PWY
GO:0000104    METACYC:SUCC-FUM-OXRED-RXN
```

+32,000 more lines...

Focusing on disagreements

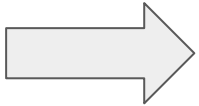
- GO:0018849 is cross-referenced to RHEA:10640
- But GO:0018849 is also mapped to EC and Metacyc terms which are cross-referenced to a different group of Rhea reactions
- ☞ GO editors examined the reactions and revised the GO:0018849 cross-reference to point to RHEA:30031



Bayesian OWL ontology merging (BOOM)

- Choose a logical interpretation for every mapping (hypothetical axioms)

```
GO:0000009   EC:2.4.1.232
GO:0000010   RHEA:20836
GO:0000016   EC:3.2.1.108
GO:0000016   METACYC:LACTASE-RXN
```



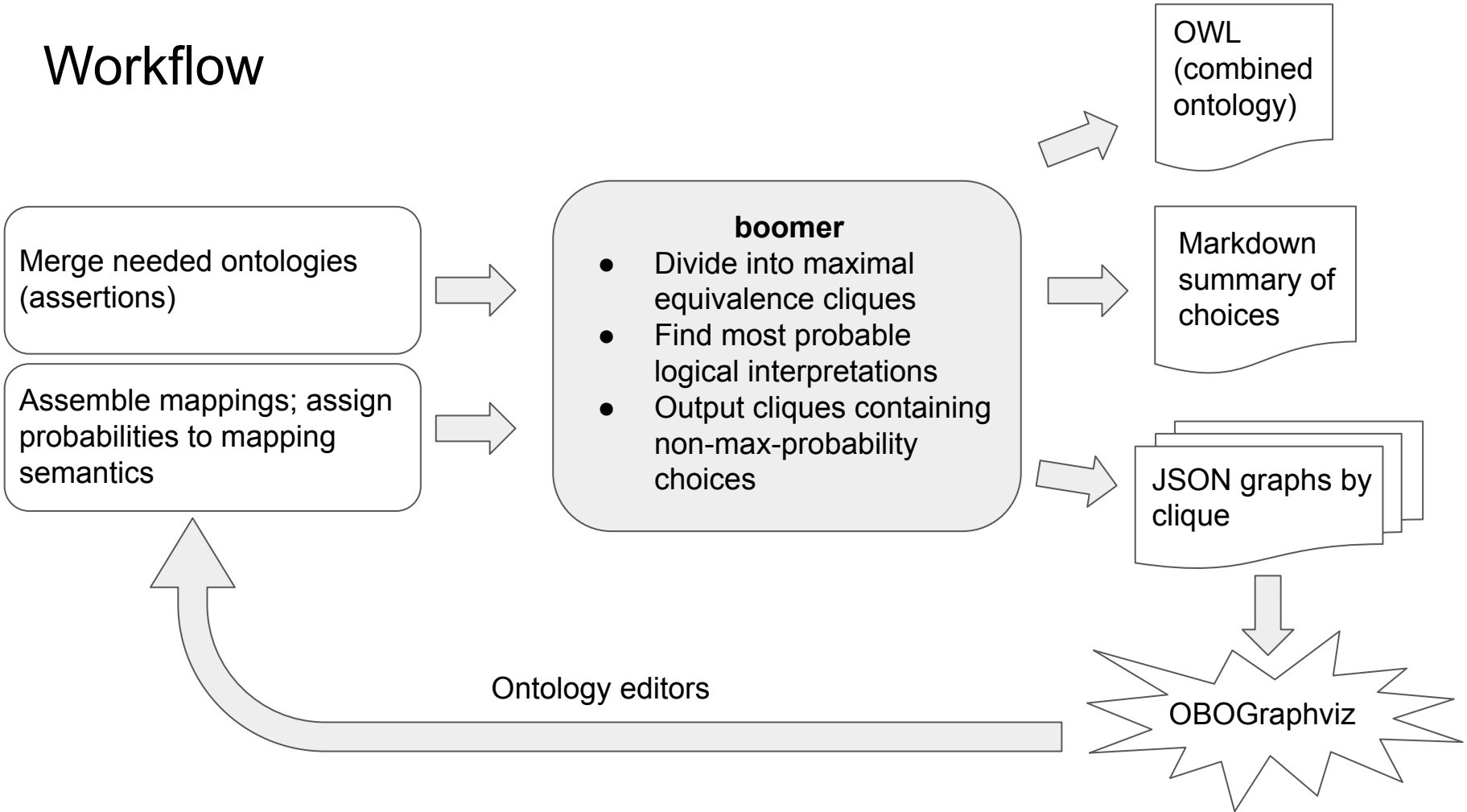
```
GO:0000009 EquivalentTo  EC:2.4.1.232
GO:0000010 SubClassOf    RHEA:20836
GO:0000016 SuperClassOf  EC:3.2.1.108
GO:0000016 <no relation> METACYC:LACTASE-RXN
```

- Combine with existing ontology (asserted axioms)
- Check the combined ontology for problems:
  - Are there now unsatisfiable classes?
    - *(e.g., from violations of disjointness)*
  - Have we inferred equivalences between terms from the same ontology?
    - *(e.g., mappings causing subclass cycles)*
  - In either case, we need to choose different logical interpretations
- Once an interpretation is chosen for every mapping, review, edit, re-run

Boomer

- k-BOOM prototype originally developed by Chris Mungall
  - Applied to initialize the Mondo harmonized disease ontology
  - Preprint: https://doi.org/10.1101/048843
- Boomer is a new implementation:
  - Simple command-line interface
  - Internally supports arbitrary sets of hypothetical axioms; use cases beyond mappings
  - Significantly faster: run in a few minutes rather than a full day
- Based on Whelk reasoner
  - Allows Boomer to quickly roll back to intermediate reasoning states
  - Whelk is implemented in Scala using immutable data structures which allows for scalable incremental reasoning

Workflow


OWL (combined ontology)
Merge needed ontologies (assertions)
boomer
Divide into maximal equivalence cliques
Find most probable logical interpretations
Output cliques containing non-max-probability choices
Markdown summary of choices
Assemble mappings; assign probabilities to mapping semantics
JSON graphs by clique
Ontology editors
OBOGraphviz

Choosing logical interpretations

Goal: find the combined set of choices *(coherent merged ontology)* with the highest joint probability

	ProperSubClassOf ⊏	ProperSuperClassOf ⊐	EquivalentTo ≡	<none> ⋢⋣
GO:0000009 EC:2.4.1.232	0.08	0.08	0.80	0.04
GO:0000010 RHEA:20836	0.08	0.08	0.80	0.04
GO:0000016 EC:3.2.1.108	0.08	0.08	0.80	0.04
GO:0000016 METACYC:LACTASE-RXN	0.08	0.08	0.80	0.04
RHEA:66549 REACTOME:R-HSA-176646.3	0.10	0.70	0.15	0.05
...and so on				

- Ahead of time: specify, for each mapping, the strength of preference for each interpretation (should add up to 1)
- Could be based on mapping relation in source:
  - **A** `skos:exactMatch` **B** ⇒ prefer **A** *EquivalentTo* **B**
  - **A** `skos:narrowMatch` **B** ⇒ prefer **A** *ProperSuperClassOf* **B**
- Or, based on lexical substring:
  - **A:**“type 1 diabetes” `xref` **B:**“diabetes” ⇒ prefer **A** *ProperSubClassOf* **B**

Searching for configurations

	Search 1				Search 2				Search 3			
	⊏	⊐	≡	⫨⫩	⊏	⊐	≡	⫨⫩	⊏	⊐	≡	⫨⫩
GO:0000009 EC:2.4.1.232	0.08	0.08	0.80	0.04	0.08	0.08	0.80	0.04	0.08	0.08	0.80	0.04
GO:0000010 RHEA:20836	0.08	0.08	0.80	0.04	0.08	0.08	0.80	0.04	0.08	0.08	0.80	0.04
GO:0000016 EC:3.2.1.108	0.08	0.08	0.80	0.04	0.08	0.08	0.80	0.04	0.08	0.08	0.80	0.04
GO:0000016 METACYC:LACTASE-RXN	0.08	0.08	0.80	0.04	0.08	0.08	0.80	0.04	0.08	0.08	0.80	0.04
RHEA:66549 REACTOME:R-HSA-176646.3	0.10	0.70	0.15	0.05	0.10	0.70	0.15	0.05	0.10	0.70	0.15	0.05
Joint probability:	0.0002048				0.0002048				0.0028672 ✅			

- Exhaustive search: check all possible configurations, from highest joint probability to lowest
  - Always finds the/a coherent merged ontology with the best possible score.
  - May never finish for large groups of mappings (15 mappings = **>1 billion** configurations)
- Greedy search: step through mappings, taking highest probability choice that fits.
  - Highly dependent on order, so conducts many runs with shuffling.

Divide and conquer

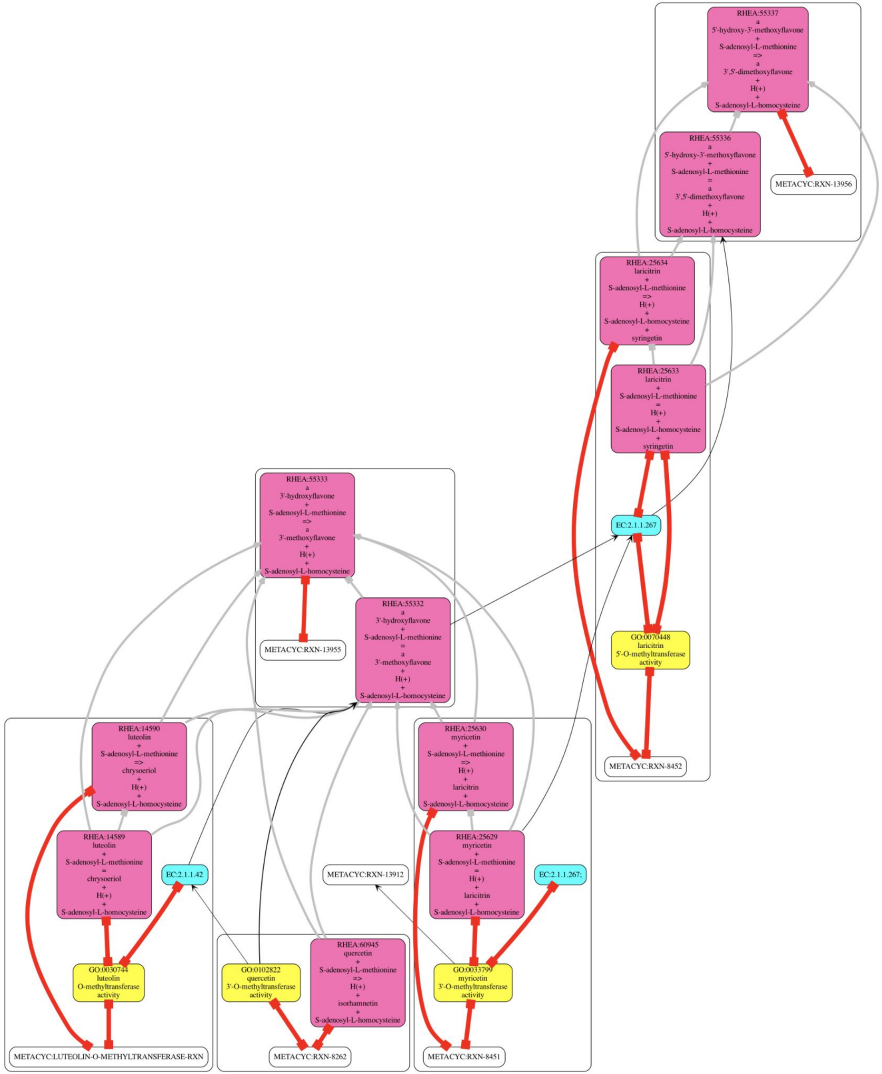
- It would be impossible to apply the exhaustive search algorithm to a full set of mappings (e.g., 32k GO–Rhea–EC-Metacyc)
- Also difficult to review resulting relation choices
- Boomer divides the input into “maximal equivalence cliques”:
  - Set all mappings to *EquivalentTo*
  - Merge with ontology and classify with reasoner (disable disjointness logic)
  - Divide mappings into groups defined by resulting equivalence nodes
  - Search for best configuration within each clique independently
- Result for each clique is output as a separate JSON graph for visualization
- *Not foolproof; resulting combination of clique results may not be coherent*
  - After deciding a logical interpretation for the mapping, differences in modeling approaches between ontologies may remain to be resolved after merging cliques.

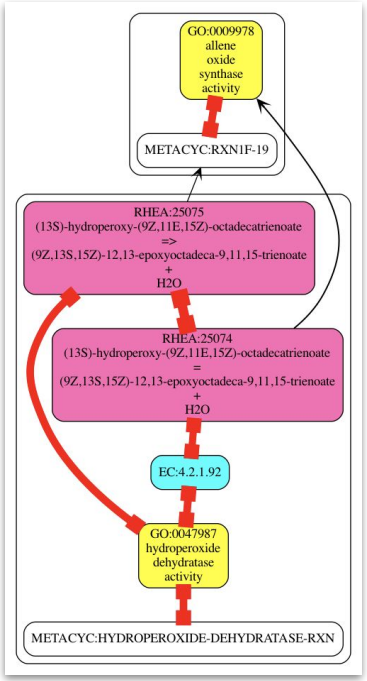
GO–Rhea mappings project

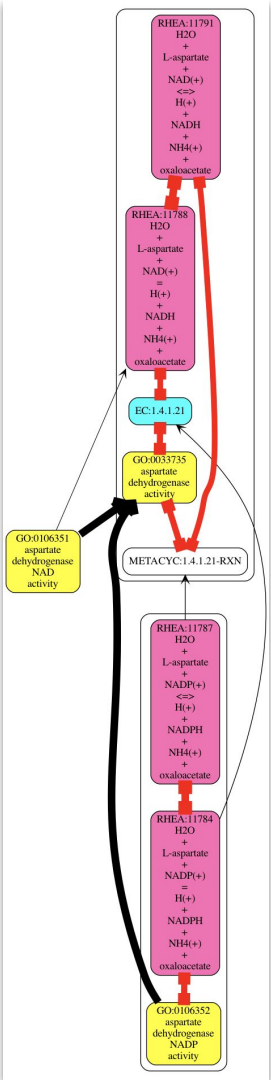
- Goal: improve the correctness of mappings between GO catalytic activity terms and Rhea reactions
  - *Enable automation of logical definitions (reaction participants) for GO catalytic activities*
- Extracted xrefs to Rhea, EC, and Metacyc from GO
- Extracted xrefs to EC and Metacyc from Rhea
- Combined xrefs and assigned probabilities; prefer *EquivalentTo* (unless GO xref has *skos:broadMatch* or *skos:narrowMatch* qualifier)
- Merged ontologies: Gene Ontology, Rhea hierarchy, EC hierarchy
- Result from Boomer: 38 cliques with a GO–Rhea mapping using a non-maximal probability
  - Issue for each, in GO GitHub repo
- Workflow Makefile

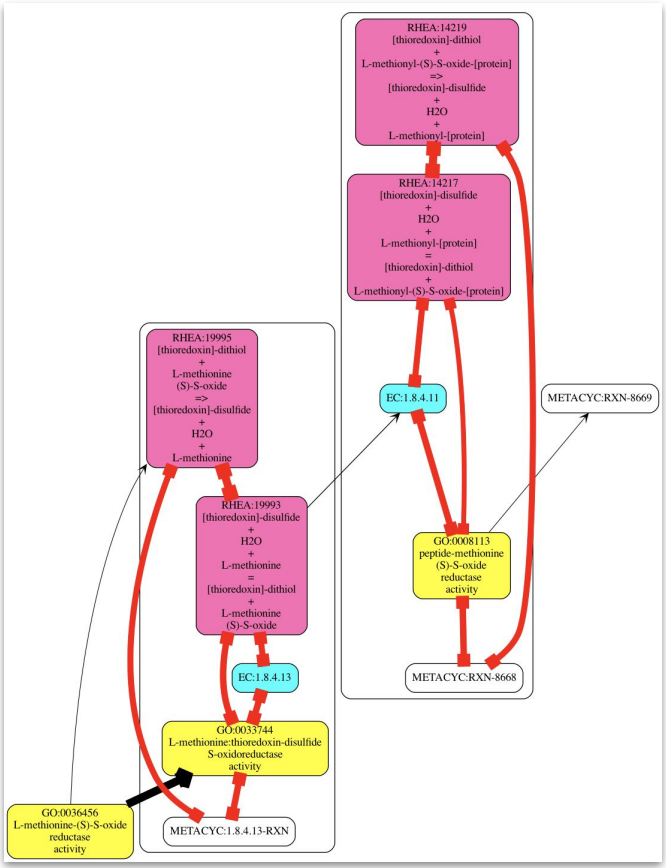
More complex example

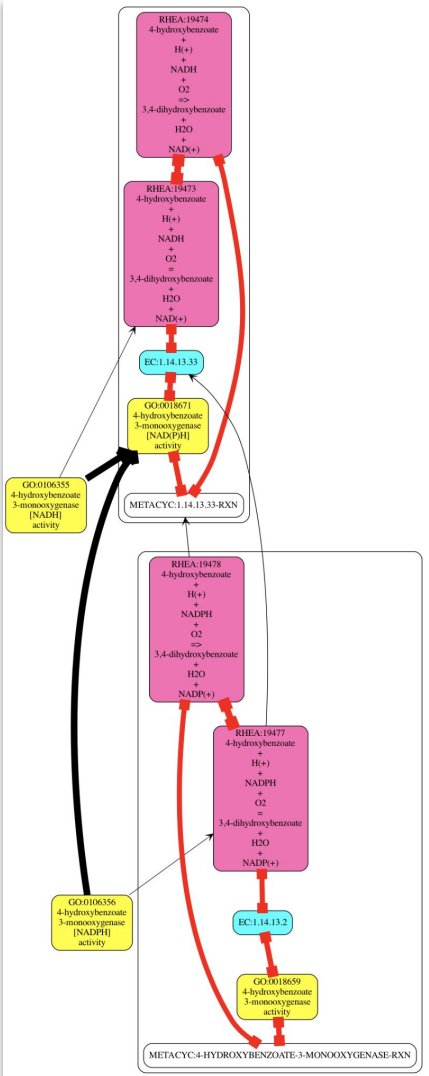
- Not always an obvious solution
- Here we can start by focusing on the cross-references around GO:0102822, which has a thin (less preferred) mapping edge
- GO:0102822 is mapped to one Rhea reaction, but grouped here with its subclass reaction.




GO:0009978
allene
oxide
synthase
activity
METACYC:RXN1F-19
RHEA:25075
(13S)-hydroperoxy-(9Z,11E,15Z)-octadecatrienoate
=>
(9Z,13S,15Z)-12,13-epoxyoctadeca-9,11,15-trienoate
+
H2O
RHEA:25074
(13S)-hydroperoxy-(9Z,11E,15Z)-octadecatrienoate
=
(9Z,13S,15Z)-12,13-epoxyoctadeca-9,11,15-trienoate
+
H2O
EC:4.2.1.92
GO:0047987
hydroperoxide
dehydratase
activity
METACYC:HYDROPEROXIDE-DEHYDRATASE-RXN


RHEA:11791
H2O
+
L-aspartate
+
NAD(+)
<=>
H(+)
+
NADH
+
NH4(+)
+
oxaloacetate
RHEA:11788
H2O
+
L-aspartate
+
NAD(+)
=
H(+)
+
NADH
+
NH4(+)
+
oxaloacetate
EC:1.4.1.21
GO:0033735
aspartate
dehydrogenase
activity
GO:0106351
aspartate
dehydrogenase
NAD
activity
METACYC:1.4.1.21-RXN
RHEA:11787
H2O
+
L-aspartate
+
NADP(+)
<=>
H(+)
+
NADPH
+
NH4(+)
+
oxaloacetate
RHEA:11784
H2O
+
L-aspartate
+
NADP(+)
=
H(+)
+
NADPH
+
NH4(+)
+
oxaloacetate
GO:0106352
aspartate
dehydrogenase
NADP
activity


RHEA:14219
[thioredoxin]-dithiol
+
L-methionyl-(S)-S-oxide-[protein]
=>
[thioredoxin]-disulfide
+
H2O
+
L-methionyl-[protein]
RHEA:14217
[thioredoxin]-disulfide
+
H2O
+
L-methionyl-[protein]
=
[thioredoxin]-dithiol
+
L-methionyl-(S)-S-oxide-[protein]
EC:1.8.4.11
GO:0008113
peptide-methionine
(S)-S-oxide
reductase
activity
METACYC:RXN-8669
METACYC:RXN-8668
RHEA:19995
[thioredoxin]-dithiol
+
L-methionine
(S)-S-oxide
=>
[thioredoxin]-disulfide
+
H2O
+
L-methionine
RHEA:19993
[thioredoxin]-disulfide
+
H2O
+
L-methionine
=
[thioredoxin]-dithiol
+
L-methionine
(S)-S-oxide
EC:1.8.4.13
GO:0033744
L-methionine:thioredoxin-disulfide
S-oxidoreductase
activity
GO:0036456
L-methionine-(S)-S-oxide
reductase
activity
METACYC:1.8.4.13-RXN


RHEA:19474
4-hydroxybenzoate
+
H(+)
+
NADH
+
O2
=>
3,4-dihydroxybenzoate
+
H2O
+
NAD(+)
RHEA:19473
4-hydroxybenzoate
+
H(+)
+
NADH
+
O2
=
3,4-dihydroxybenzoate
+
H2O
+
NAD(+)
EC:1.14.13.33
GO:0018671
4-hydroxybenzoate
3-monooxygenase
[NAD(P)H]
activity
GO:0106355
4-hydroxybenzoate
3-monooxygenase
[NADH]
activity
METACYC:1.14.13.33-RXN
RHEA:19478
4-hydroxybenzoate
+
H(+)
+
NADPH
+
O2
=>
3,4-dihydroxybenzoate
+
H2O
+
NADP(+)
RHEA:19477
4-hydroxybenzoate
+
H(+)
+
NADPH
+
O2
=
3,4-dihydroxybenzoate
+
H2O
+
NADP(+)
GO:0106356
4-hydroxybenzoate
3-monooxygenase
[NADPH]
activity
EC:1.14.13.2
GO:0018659
4-hydroxybenzoate
3-monooxygenase
activity
METACYC:4-HYDROXYBENZOATE-3-MONOOXYGENASE-RXN

Future plans for Boomer

- Refine workflow with additional cases
  - Some example workspaces (poorly documented!):
    - GO + Rhea
    - Biolink model relations + Relation Ontology
    - UBERON + EMAPA anatomy
- Explore options for more complex modularization strategies
  - Balance modularization with preference for global coherence
- Applications beyond mapping
  - Boomer can be used to find compatible combinations of any kind of class axioms (subclass, equivalence, disjointness, class and property assertions)
- Command-line interface continues to evolve
  - Plan to add a separate command to output a fully coherent merged ontology (as an alternative to current clique approach)
- Support SSSOM inputs

Thank you!

Software:

- **Boomer: https://github.com/INCATools/boomer**
- OBOGraphviz: https://github.com/cmungall/obographviz
- ROBOT: http://robot.obolibrary.org
  - *talk by Rebecca Jackson tomorrow*
- Whelk reasoner: https://github.com/balhoff/whelk

(All open source)

Funding:

- NIH 1U01HG009453-01 (INCA)
- NIH 2U41HG002273-17 (Gene Ontology Consortium)

Special thanks to Pascale Gaudet (GO), Harold Drabkin (GO), Anne Morgat (Rhea)