Cross-species comparison of GO annotations : advantages and limitations of semantic similarity measures

O. Dameron, C. Bettembourg, L. Joret

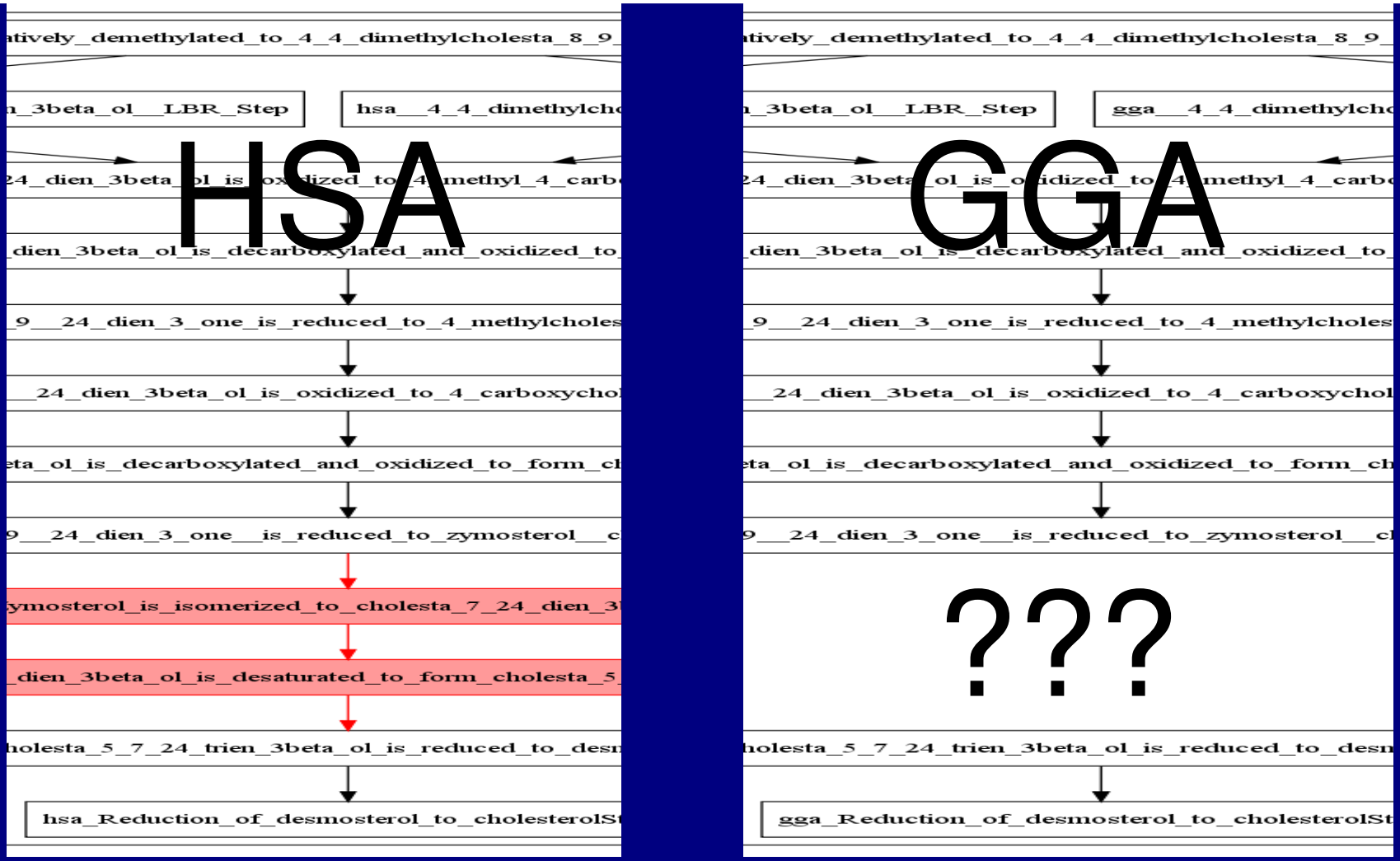
U936 “Conceptual modeling of biomedical knowledge”
Université de Rennes 1, France
http://www.u936.univ-rennes1.fr

Context: NAFLD

- Fatty Liver Disease = lipid infiltration in liver parenchyma cells
- Non-alcoholic fatty liver disease:
  - 6% to 24% of worldwide population

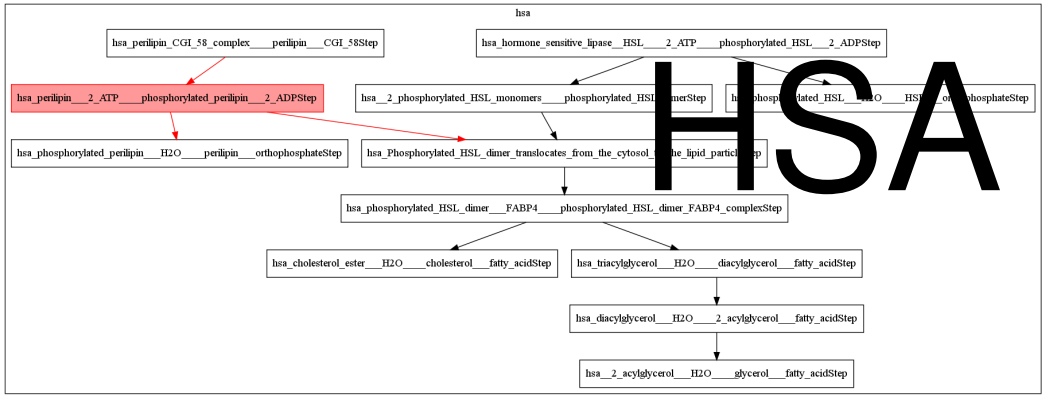
    USA: 1/3 adults et 1/10 children+teenagers
  - Increased prevalence if overweight or obesity
  - Evolution: NASH, fibrosis, cirrhosis, hepatocellular carcinoma
- lipid metabolism conserved among sup eukaryots
  - But chicken seem more resistant to liver cirrhosis

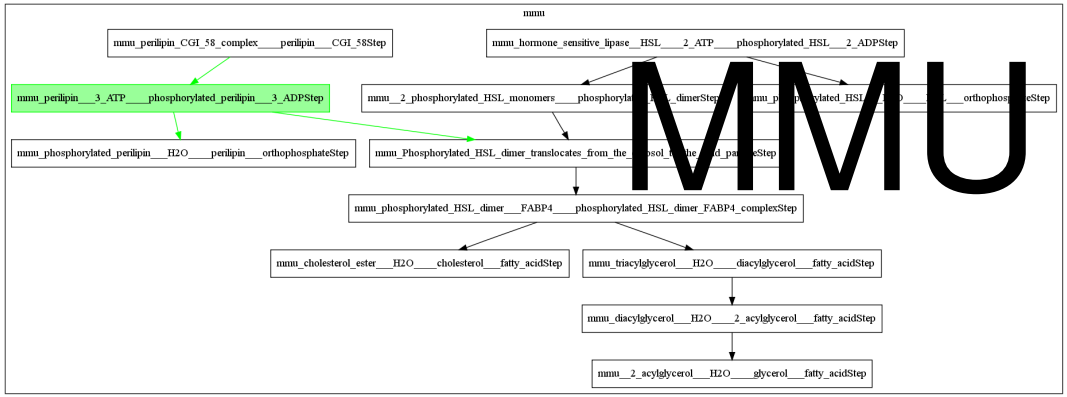
Transformation of lanosterol to cholesterol (HSA-GGA)



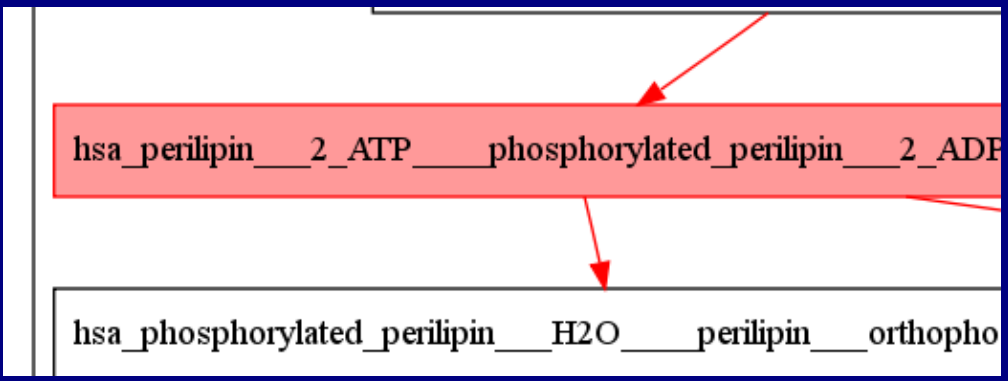
- Some steps seem species-specific (here HSA)
  - We **do not know** if they exist for the other species

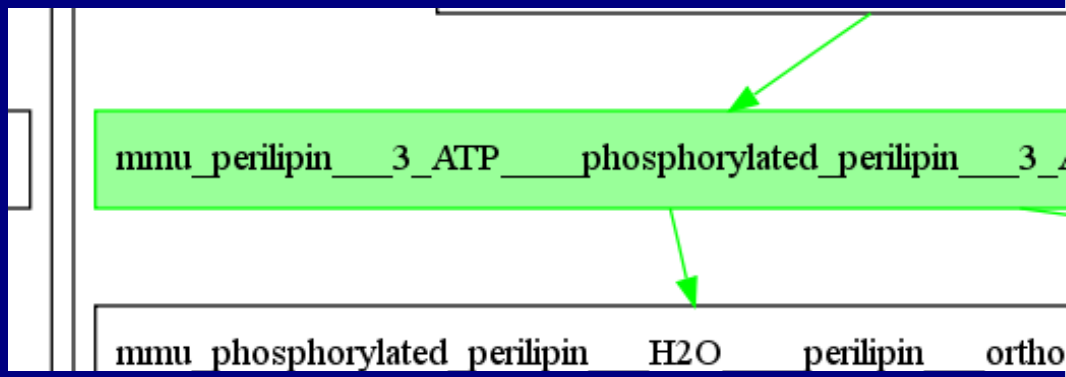
How different different pathway steps really are?





Hormone_sensitive_lipase__HSL__mediated_triacylglycerol_hydrolysis





Hormone sensitive lipase HSL mediated triacylglycerol hydrolysis (HSA - MMU)

Hypothesis

Compare the GO annotations of the gene products involved in each pathway step

- Measure overlap and specificities
  - Granularity can be addressed with GO hierarchy
- Detect difference in annotations of otherwise perfectly homologous steps

Approach

- Cross-species comparison of 1 gene product annotations
  - Validate on Apoa1 (known to be different) and Apoa5 (known to be similar) for HSA and MMU
- Generalize to compare annotations of sets of gene products involved in 1 pathway step

Material and methods

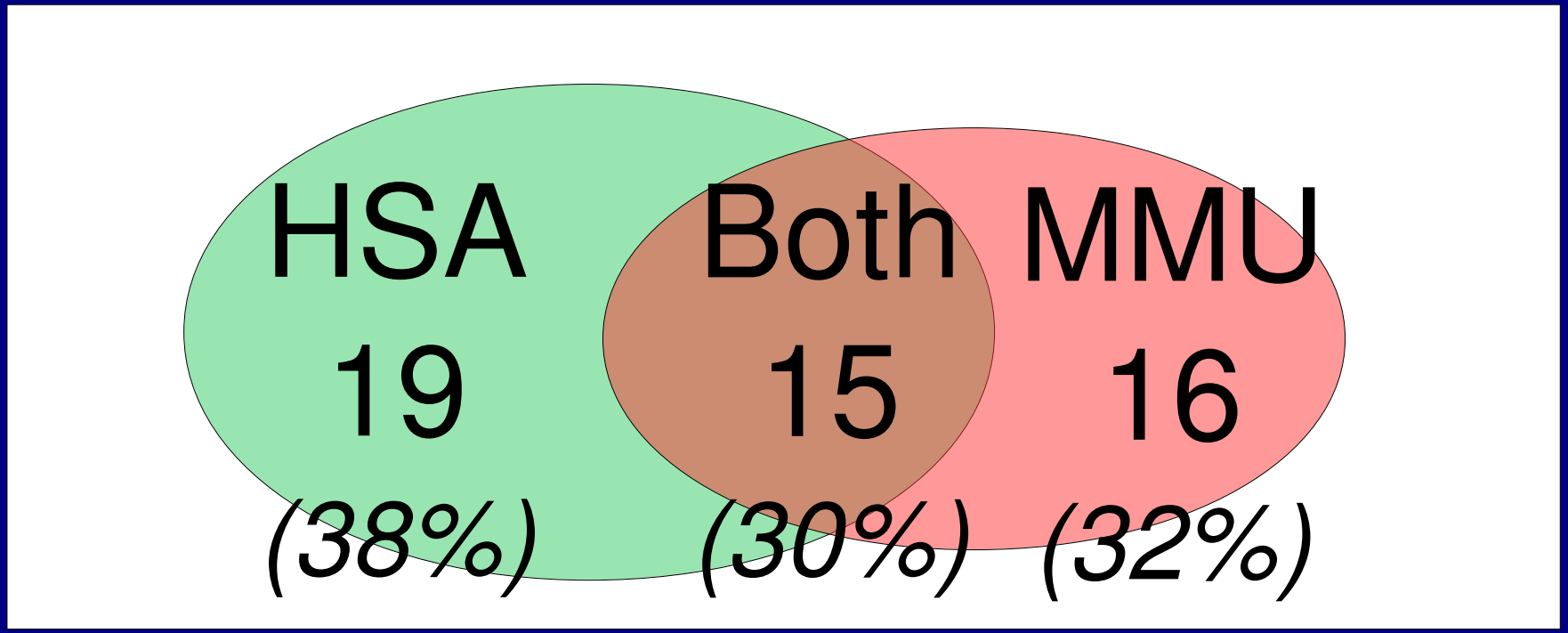
- Retrieve GO annotations from EBI GOA database for each species (*H. Sapiens* and *Mus Musculus*)
- Compare the two sets of annotations
  - Identify limitations of straightforward approach
  - Use Wang's semantic similarity measure
- Apply to
  - Apoa1 (which we know is different btw HSA and MMU
  - Apoa5 (which we know is similar btw HSA and MMU

Using set cardinality to compare two sets of GO annotations

(after possible filtering or enriching)

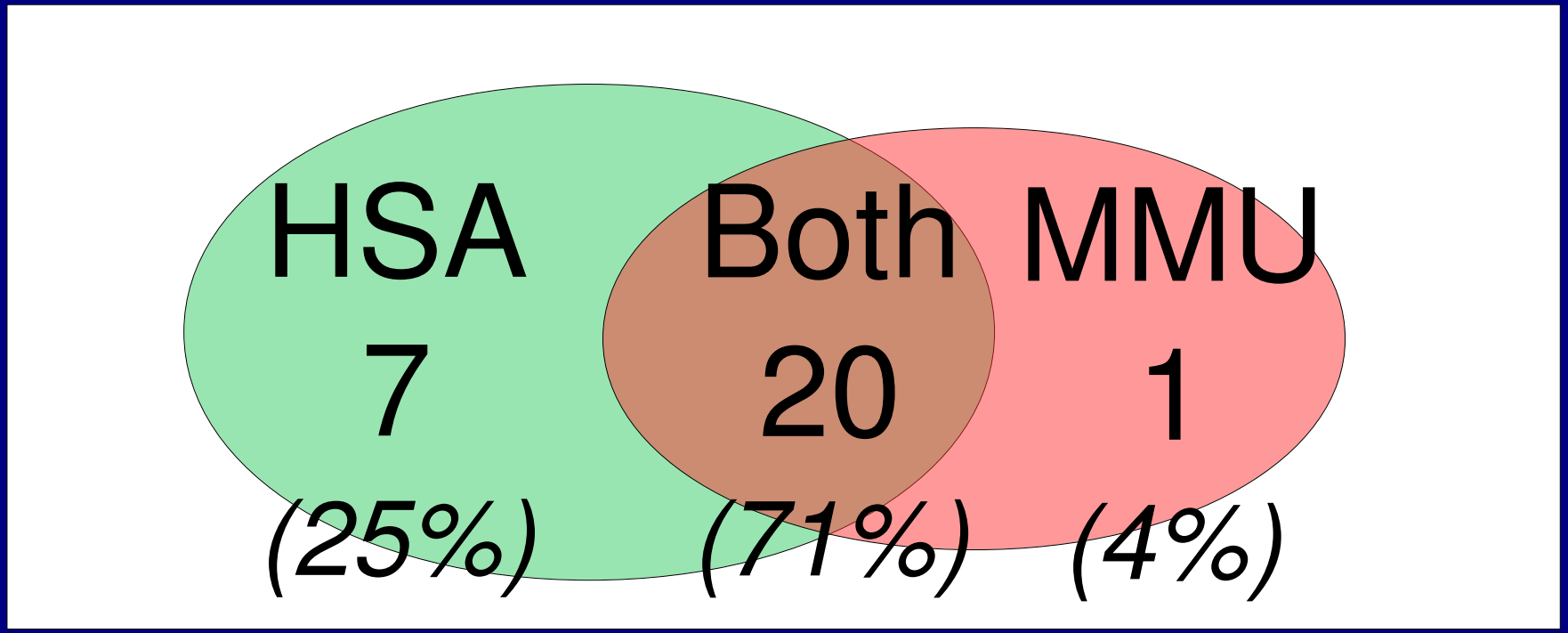
Results: APOA1 hsa/mmu

- Raw comparison (EBI GOA database)
- HSA: 34
- MMU: 31

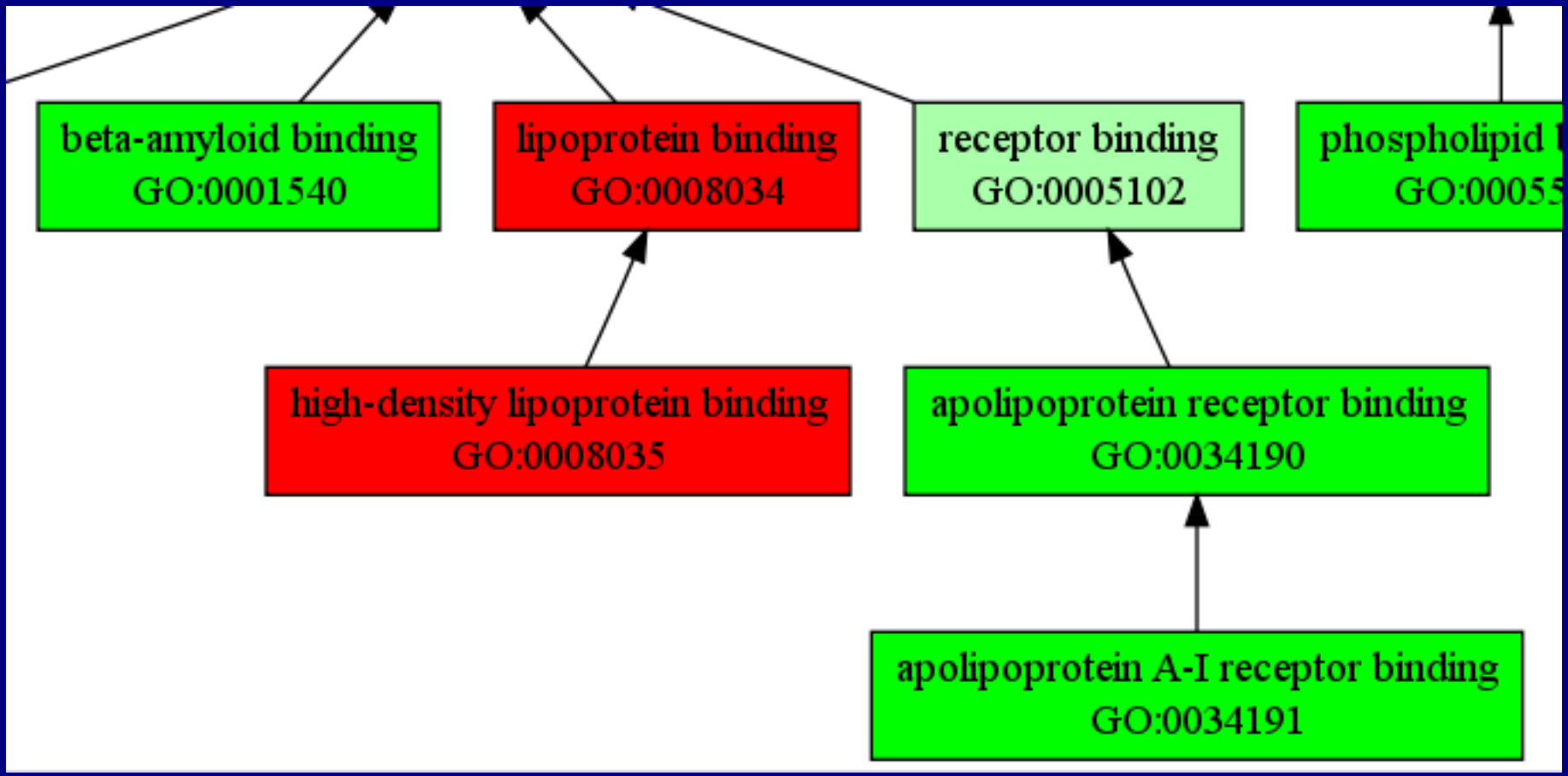


Results: APOA5 hsa/mmu

- Raw comparison (EBI GOA database)
- HSA: 27
- MMU: 21



Problem 1: redundant annotations

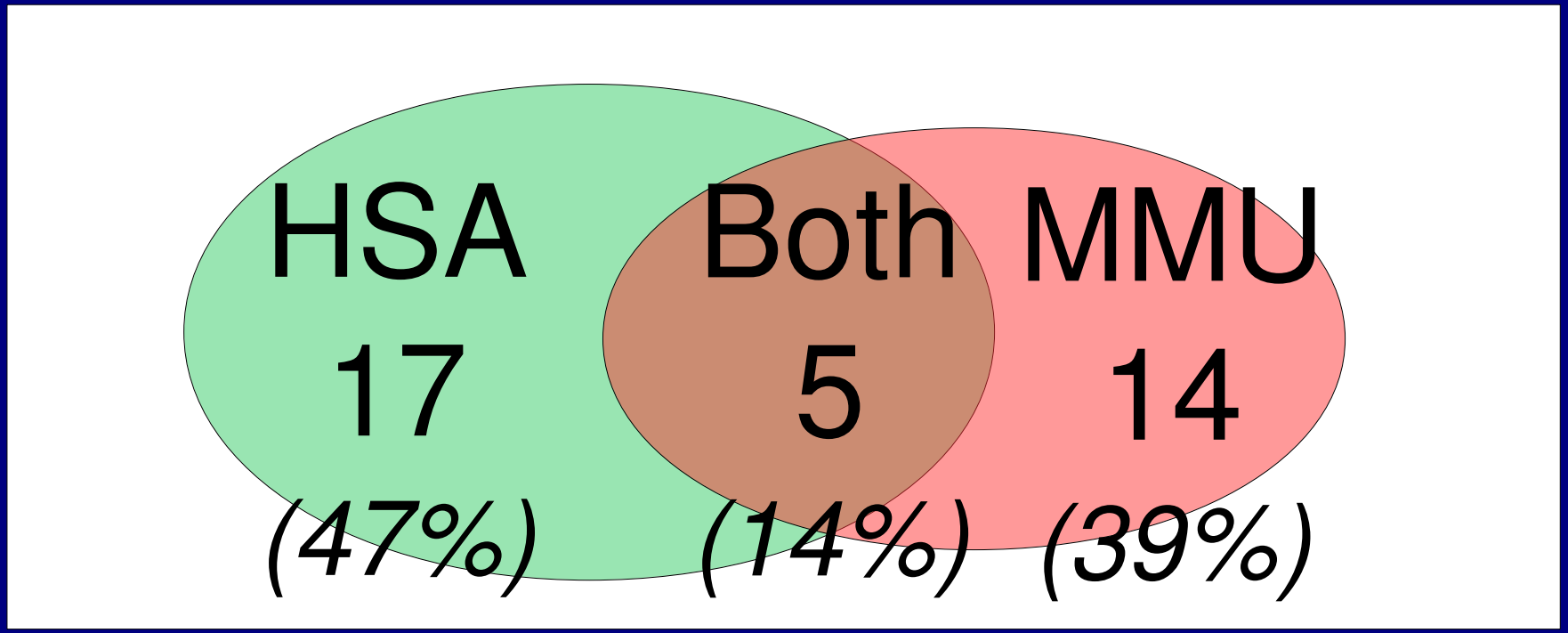


Redundancy favoring
MMU specificity

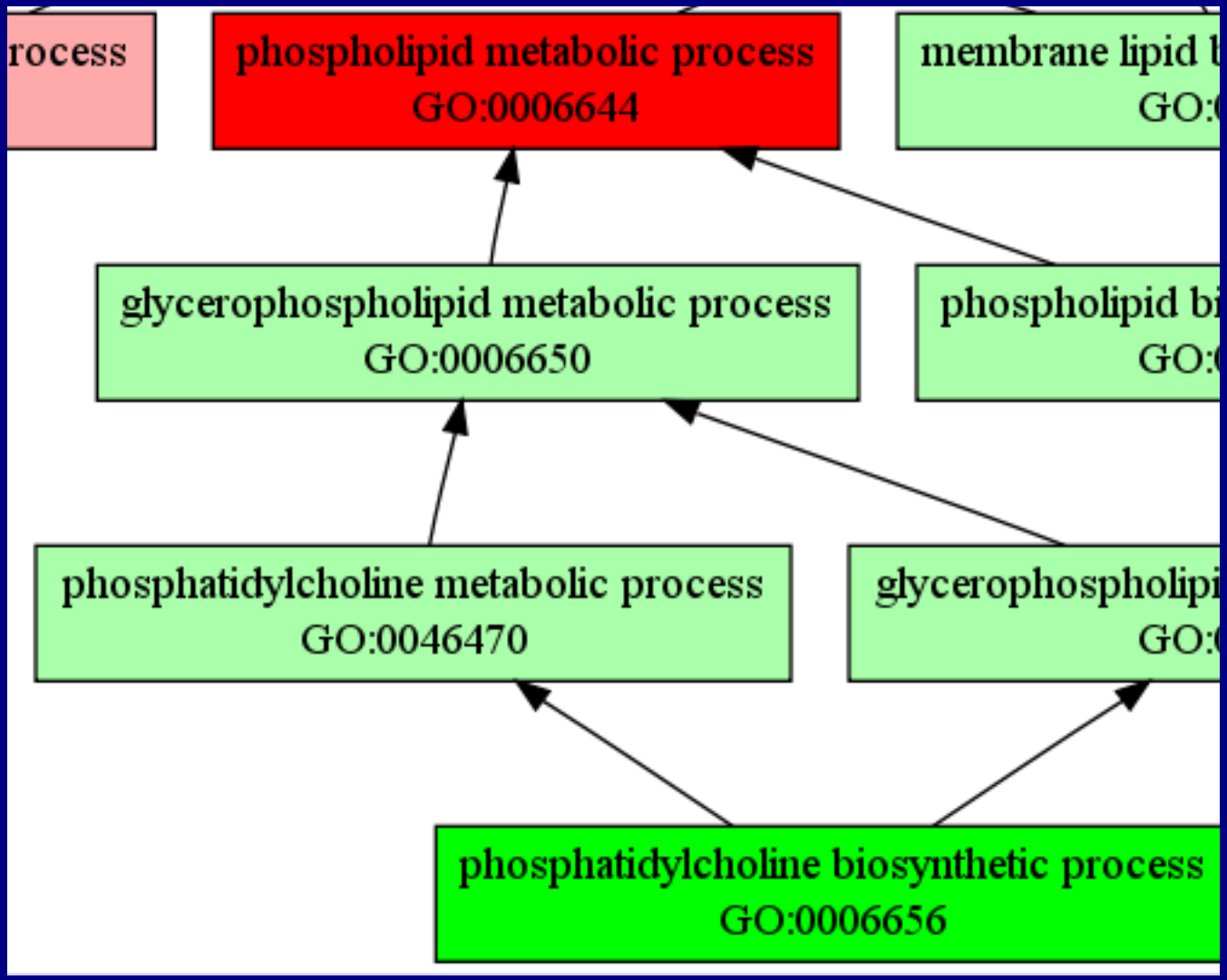
Redundancy favoring
HSA specificity

Considering only leaves

- Leaves (EBI GOA database) : Apoa1
- HSA: 21 (was 34)
- MMU: 19 (was 31)



Problem 2: annotations with different granularities



MMU-specific annotation (according to true path rule, it should be counted as common)

HSA-specific annotation

Problem 2: annotations with different granularities

- BUT, some annotations have different granularities, which introduces a bias
- Solution: for each species, retrieve all the ancestors of the annotations and compute specificity on these expanded sets
  - Bonus: the redundancy problem disappears

Ancestors: APOA1 hsa/mmu

- Expanded to ancestors (EBI GOA database)
- HSA: 117
- MMU: 104

	HSA		Common		MMU	
Initial data	19	*38.00%*	15	*30.00%*	16	*32.00%*
Leaves	17	*47.22%*	5	*13.89%*	14	*38.89%*
Expanded	76	*42.22%*	41	*22.78%*	63	*35.00%*

- Note the evolution of %

Problem 3: negation

- Not finding an annotation for one species only means “we do not know whether the annotation is valid for this species or not”
- GOA supports the NOT modifier for representing “we know that this annotation is not true”
- We know that for MMU, Apoa1 is not associated with:
  - “axon regenation” (GO:0031103)
  - “protein localization” (GO:0008104)
- These should be counted too, but separately

Results: APOA1 hsa/mmu

- Expanded to ancestors (EBI GOA database)
- HSA: 117
- MMU: 104

		HSA		Common		MMU	
Initial data	positive	19	*39.58%*	15	*31.25%*	14	*29.17%*
	negative	0	*0.00%*	0	*0.00%*	2	*100.00%*
	Non diff.	19	*38.00%*	15	*30.00%*	16	*32.00%*
Leaves	positive	17	*50.00%*	5	*14.71%*	12	*35.29%*
	negative	0	*0.00%*	0	*0.00%*	2	*100.00%*
	Non diff.	17	*47.22%*	5	*13.89%*	14	*38.89%*
Expanded	positive	76	*48.10%*	41	*25.95%*	41	*25.95%*
	negative	0	*0.00%*	0	*0.00%*	22	*100.00%*
	Non diff.	76	*42.22%*	41	*22.78%*	63	*35.00%*

Results: APOA5 hsa/mmu

- Expanded to ancestors (EBI GOA database)
- HSA: 118
- MMU: 93

		HSA		Common		MMU	
Initial data	positive	6	*22.22%*	20	*74.07%*	1	*3.70%*
	negative	1	*100.00%*	0	*0.00%*	0	*0.00%*
	Non diff.	7	*25.00%*	20	*71.43%*	1	*3.57%*
Leaves	positive	5	*25.00%*	15	*75.00%*	0	*0.00%*
	negative	1	*100.00%*	0	*0.00%*	0	*0.00%*
	Non diff.	6	*28.57%*	15	*71.43%*	0	*0.00%*
Expanded	positive	20	*17.70%*	93	*82.30%*	0	*0.00%*
	negative	5	*100.00%*	0	*0.00%*	0	*0.00%*
	Non diff.	25	*21.19%*	93	*78.81%*	0	*0.00%*

Synthesis

- GO semantics must be taken into account (not a surprise!)
  - Redundancy
  - Differences of granularity
  - Negation
- Preprocessing (filtering and enriching) introduces a new bias artificially promoting common annotations
- Need for finer comparison technics

Using semantic similarity to compare two sets of GO annotations

GO-specific semantic similarity (Wang)

Semantic similarity between 2 concepts C1 and C2: sum of the semantic contribution of all ancestors common to C1 and C2, divided by the semantic values of C1 and of C2

- GO term A is represented by $DAG_A = (A, T_A, E_A)$
  - T_A: A and all its ancestors (is_a or part_of)
  - E_A: set of relations connecting elts in T_A

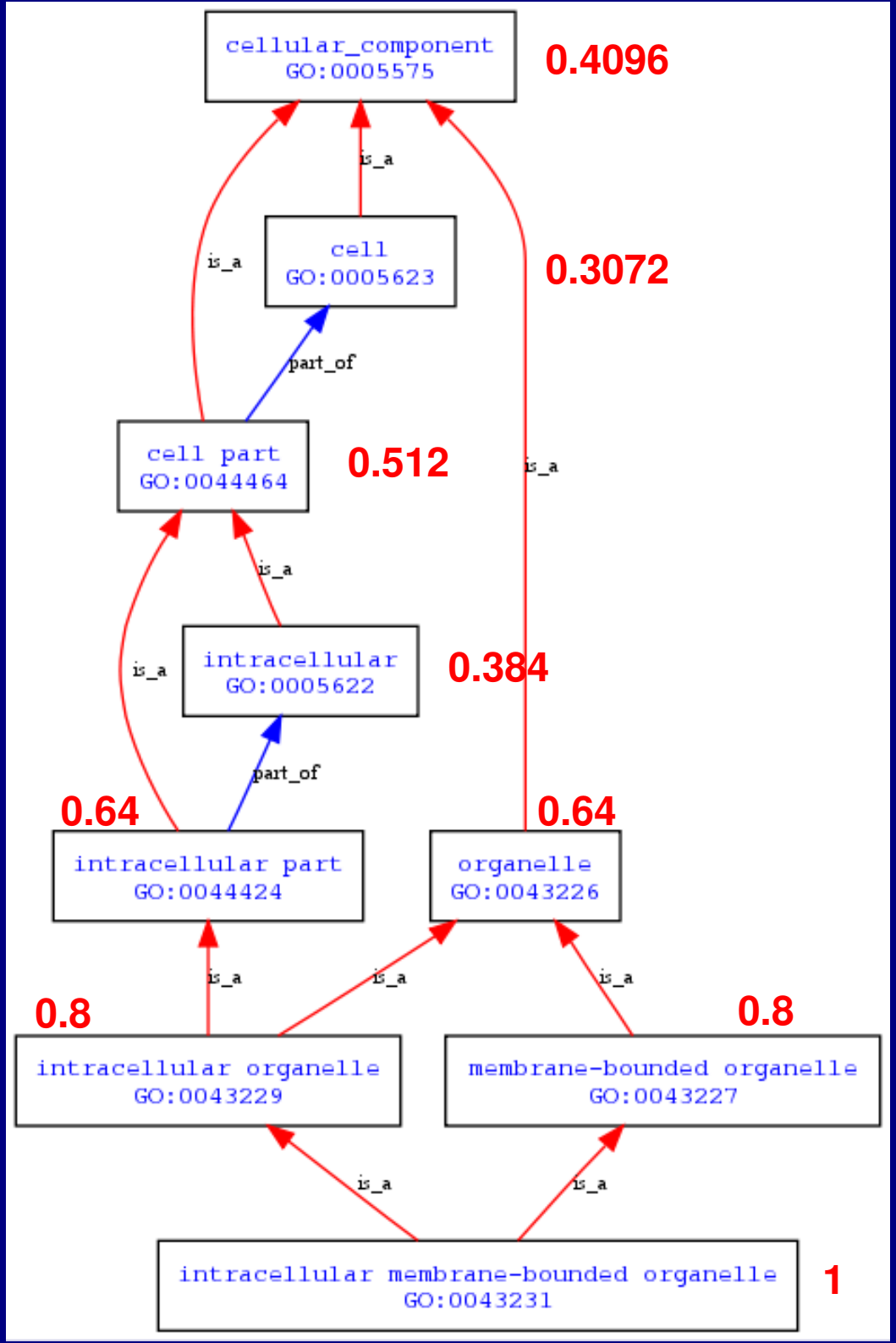
Contribution of term t to the semantics of term A

- $S_A(A) = 1$
- $S_A(t) = \max_{t' \in \text{children of } t} w * S_A(t')$

W: weight of the relation between t' and t (proposed experimentally by Wang et al.)

- is_a: 0.8
- part_of: 0.6

Semantic contributions of ancestors to GO:0043231



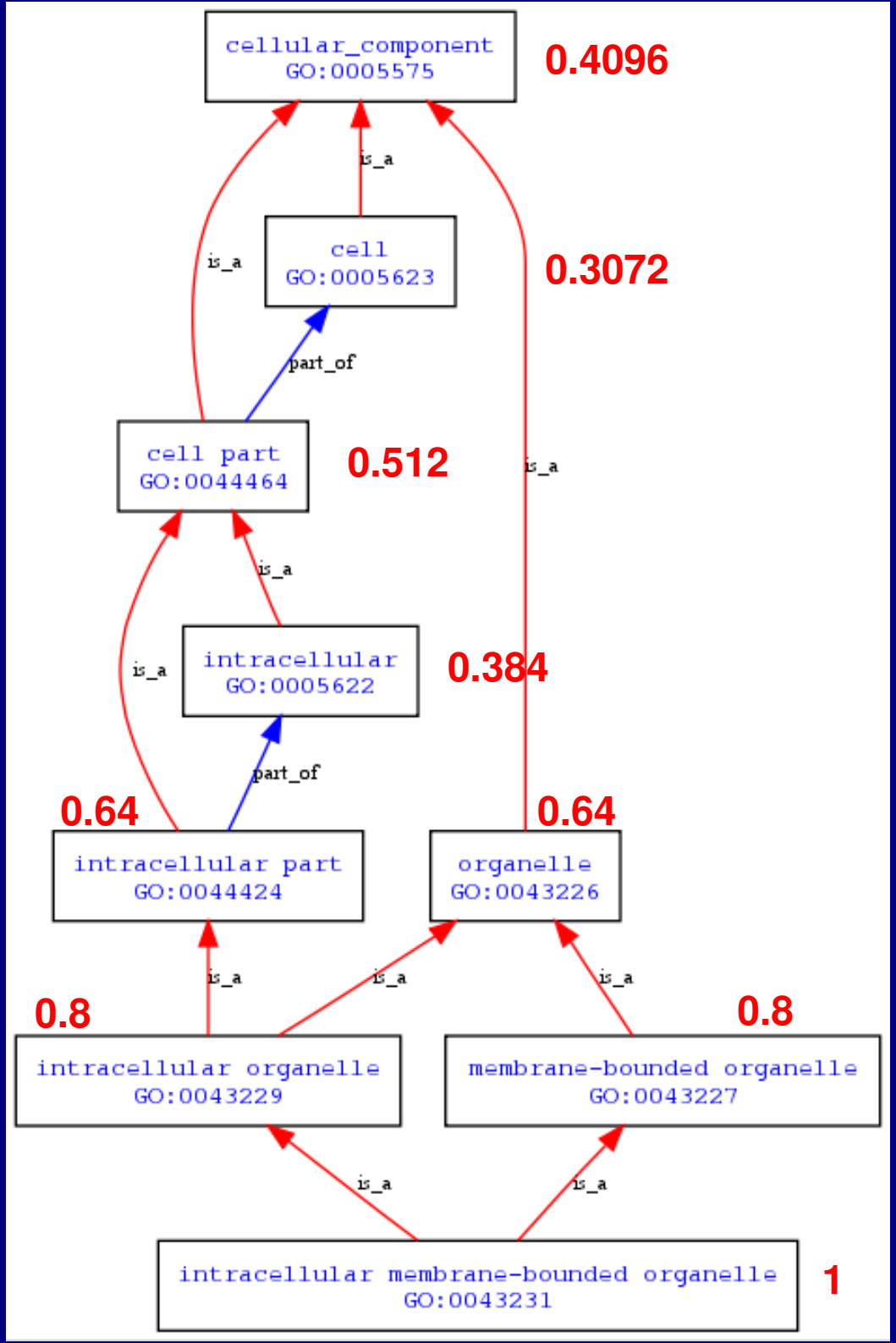
- Terms closer to GO:0043231 contribute more
- The farther the ancestor, the smaller its contribution

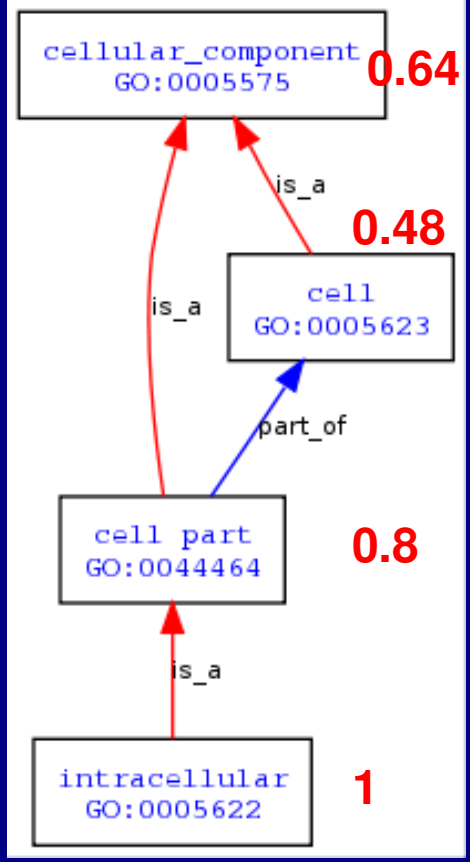
Semantic value of a term

$$SV(A) = \sum_{t \in T_A} SA(t)$$

The semantic value of a term A is the sum of the semantic contributions of all its ancestors

In the previous example $SV_{GO:0043231} = 5.5952$





SV(GO:0005622) = 2.92

The more general a term, the smaller its semantic value

SV(GO:0043231) = 5.5952

Semantic similarity of 2 terms

$$S_{GO}(A,B) = \frac{\sum_{t \in T_A \cap T_B} (SA(t) + SB(t))}{SV(A) + SV(B)}$$

$\forall (A,B),\ S_{GO}(A,B) \in [0;1]$

Example: $S_{GO}(0043231;0043229) = 0.7727$

Semantic similarity of term t and set of terms A

$$\mathbf{Sim(t,A) = \max_{a \in A} S_{GO}(t,a)}$$

The semantic similarity between a term *t* and a set of terms A is the semantic similarity of *t* and its closest element in A

Semantic similarity of 2 sets of terms

$$\mathrm{Sim}(A,B) = \frac{\sum_{1 \le i \le m} \mathrm{Sim}(a_i,B) + \sum_{1 \le j \le n} \mathrm{Sim}(b_j,A)}{m + n}$$

Wang semantic similarity of apoa1 between hsa and mmu

- Apoa1: 0.719393
- Apoa5: 0.957423

Contrary to assertions in Wang et al.'s article, we found from analysis of several example that the limit between similar sets and dissimilar sets is not 0.5, but rather somewhere between 0.7 and 0.8

See limitation #5 in a few slides

Limits of Wang semantic similarity (1/6)

- Negation is ignored
  - Easy: remove negated annotations from the set
  - Better : differentiate
    - not(GO:xxxxxx) for species1 and ??? for species2
    - not(GO:xxxxxx) for species1 and GO:xxxxxx for sp2
    - not(GO:xxxxxx) for sp1 and not(GO:xxxxxx) for sp2

Limits of Wang semantic similarity (2/6)

- Evidence codes are ignored
  - Should be processed between annotations retrieval and semantic similarity computation?
  - Should be exploited by semantic similarity?

Limits of Wang semantic similarity (4/6)

- Should be computed separately for BP, CC, MF

Computing semantic similarity separately on BP, CC and MF

- Previous example about GO:004323 not relevant (all annotations are cellular component-related)
- apoa1 / apoa5:

	Apoa1	**Apoa5**
GO	0.6579	0.9367
BP	0.6039	0.9248
CC	0.5229	0.9039
MF	0.8213	0.9689

Limits of Wang semantic similarity (5/6)

- Redundancy is still an issue
  - Should be computed on leaves
- Difference of granularities is addressed

Redundancy-robust semantic similarity of sets of annotations

$$Sim(A,B) = \frac{\sum_{1 \le i \le m} Sim(a_i,B) + \sum_{1 \le j \le n} Sim(b_j,A)}{m + n}$$

$$Sim(t,A) = \max_{a \in A} S_{GO}(t,a)$$

$$S_{GO}(a,b) = \frac{\sum_{t \in T_a \cap T_b} (S_a(t) + S_b(t))}{SV(a) + SV(a)}$$

Redundancy-robust semantic similarity of sets of annotations

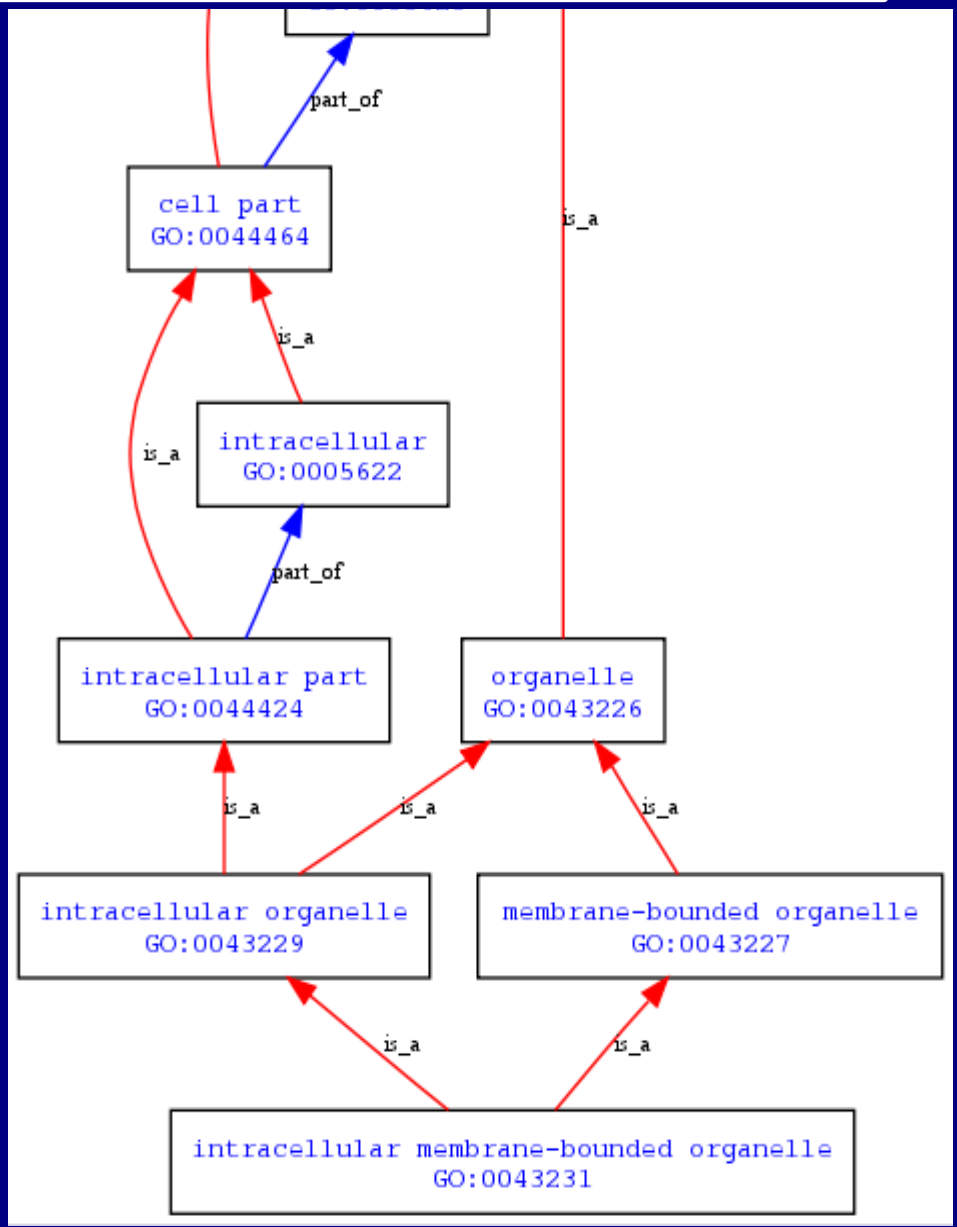
- apoa1 / apoa5:

	Apoa1			Apoa5		
	Initial	***Leaves***	***Ancestors***	***Initial***	***Leaves***	***Ancestors***
GO	0.6579	0.4787	0.7544	0.9367	0.9025	0.9412
BP	0.6039	0.3754	0.7664	0.9248	0.8467	0.9485
CC	0.5229	0.5849	0.5354	0.9039	0.9039	0.8207
MF	0.8213	0.6564	0.8724	0.9689	0.9659	0.9957

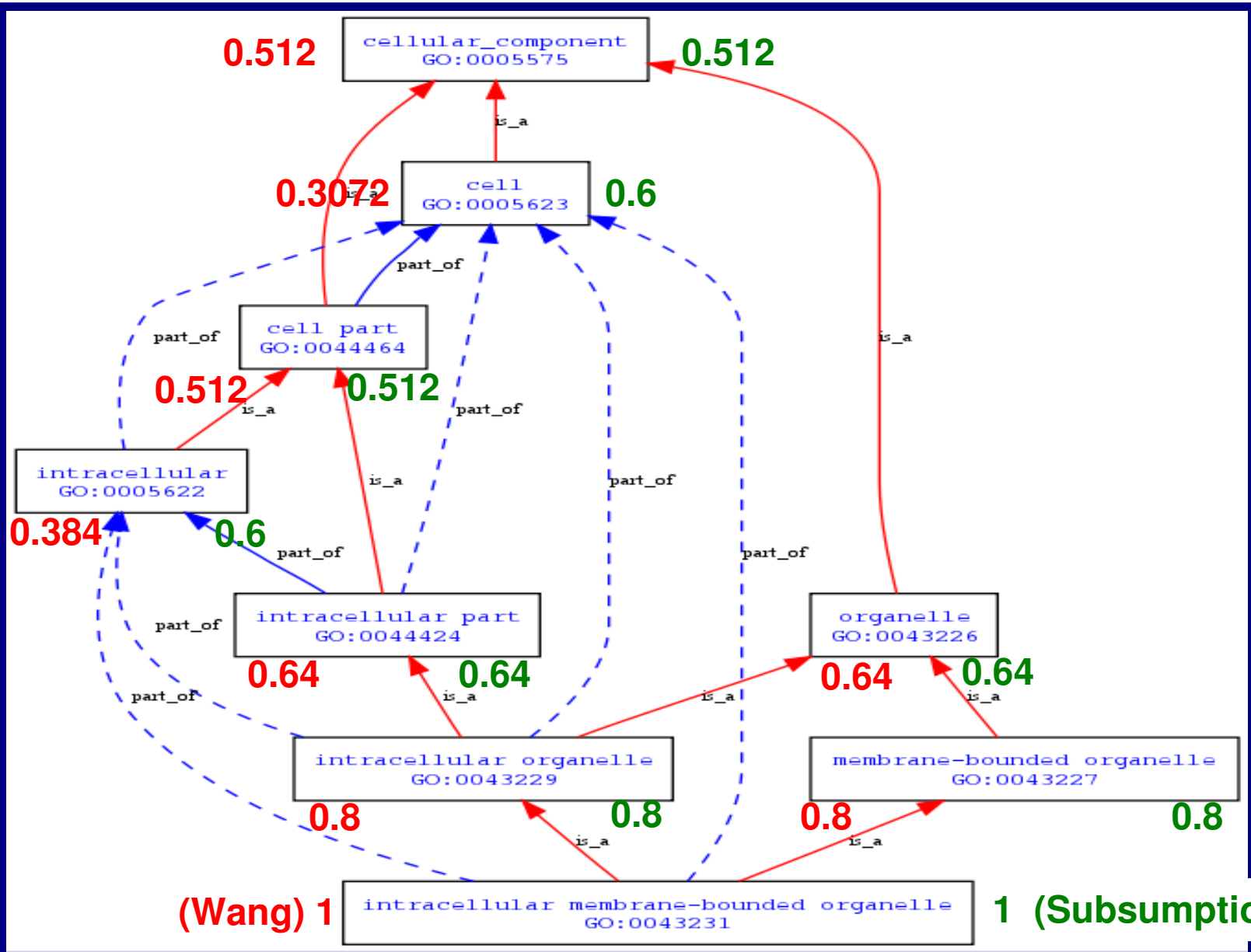
- Initial data probably contain redundancies; ancestors-enriched certainly do!
- This introduces a bias
- Compare only the more specific annotations

Limits of Wang semantic similarity (6/6)

- Inheritance is ignored
  what kind of
  “semantic” similarity
  is this? :-)



Subsumption-compliant semantic similarity



Subsumption-compliant semantic similarity: results

- Semantic value of GO:0043231
  - Wang: 5.5952
  - Subsumption-compliant: 6.1040

Subsumption-compliant semantic similarity: apoa1

- Initial data
  - Wang: 0.7194
  - Subsumption-compliant: 0.7207
- Leaves
  - Wang: 0.5050
  - Subsumption-compliant: 0.5097
- Ontology structure analysis:
  - hsa: 1643 is_a 73 part_of
  - mmu: 1476 is_a 27 part_of

Subsumption-compliant semantic similarity: apoa5

- Initial data
  - Wang: 0.9574
  - Subsumption-compliant: 0.9584
- Leaves
  - Wang: 0.9176
  - Subsumption-compliant: 0.9189
- Ontology structure analysis:
  - hsa: 805 is_a 33 part_of
  - mmu: 559 is_a 15 part_of

Subsumption-compliant semantic similarity: conclusion

- Theoretically important
- Practically, the differences are small :-(
- But:
  - # is_a >> # part_of
  - The (few) part_of relations are not uniformly distributed among BP, CC and MF
  - The structure of GO may also introduce a bias (terms such as “Intracellular part” or “Cell part” promote is_a)

Conclusion

Conclusion

Semantic comparison of sets of GO annotations

- Missing annotation data is a serious limitation
- The semantics of the annotations has to be considered
- Different strategies for comparing
  - Set overlap and set difference
  - Wang semantic similarity
- All fail to fully leverage the (fortunately limited) semantics of GO