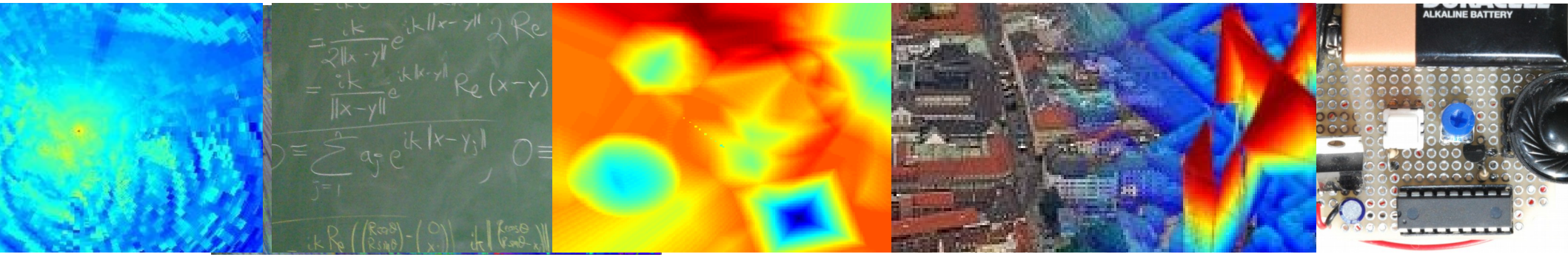
Finding cross-species orthologs with local topology



Michael Robinson

Acknowledgements

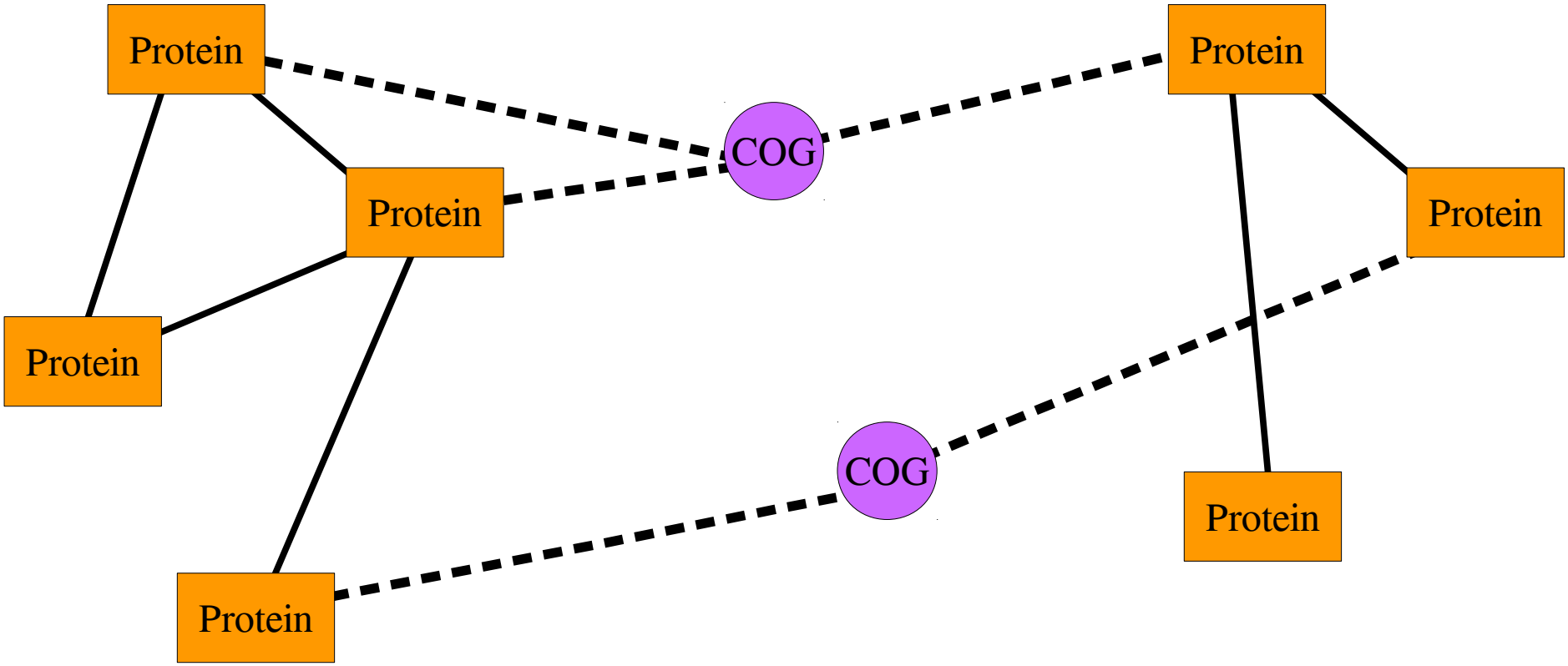
- SRC: Chris Capraro
- PNNL:
  - Cliff Joslyn
  - Katy Nowak
  - Brenda Praggastis
  - Emilie Purvine
- Baylor College of Medicine:
  - Olivier Lichtarge
  - Angela Wilkins
  - Daniel Konecki
- Reza Ghanadan (DARPA/DSO SIMPLEX program)

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

Problem statement

- Can we identify related proteins across species?

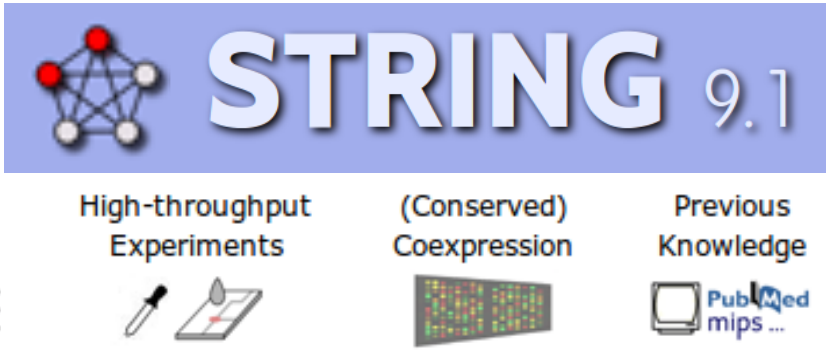


COG = Clusters of Orthologous Groups – set of genetically related proteins

Working dataset

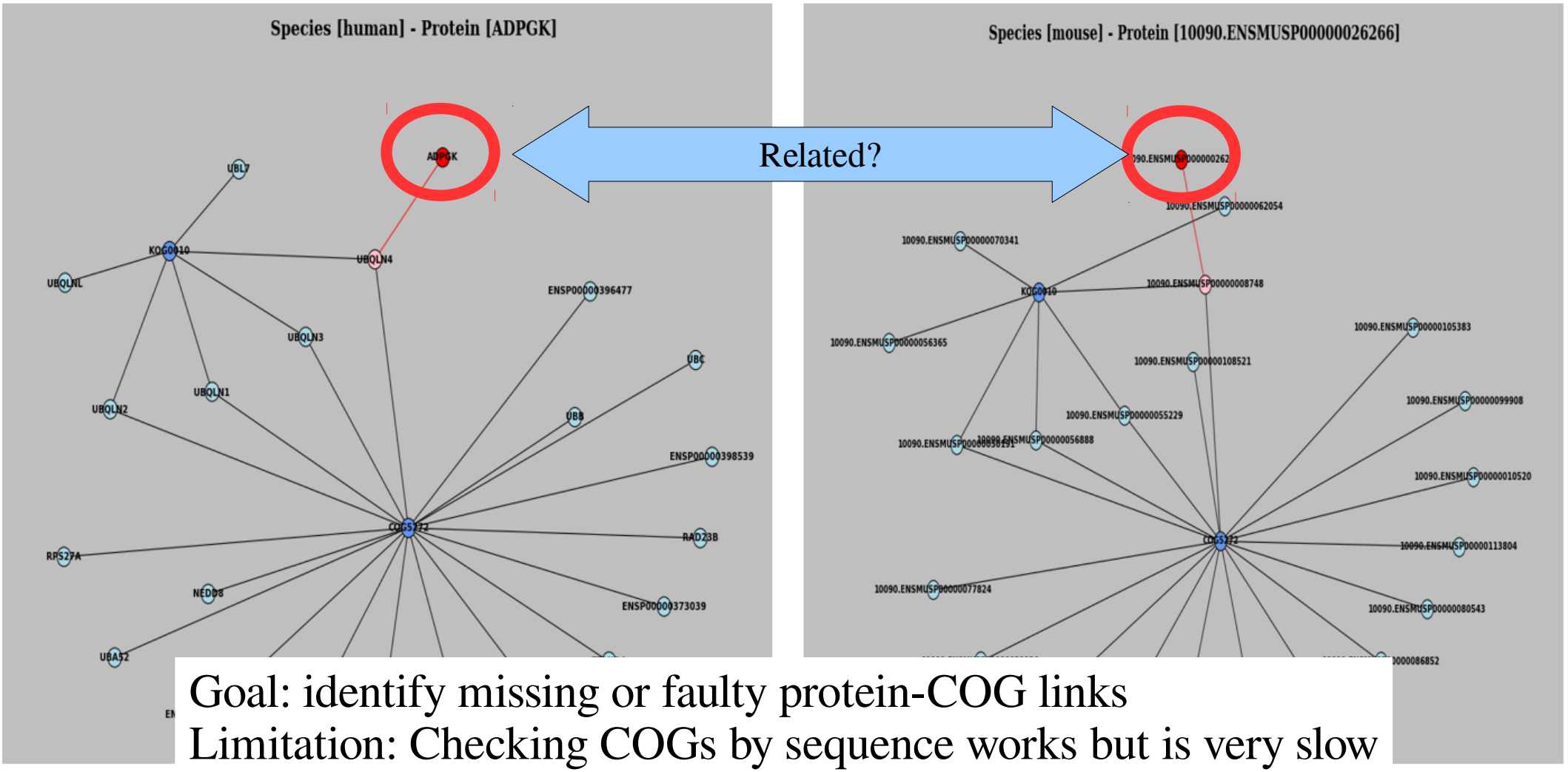
- Source: StringDB version 9.1

  http://string91.embl.de/



  - Protein-protein interactions
  - Clusters of Orthologous Groups (COGs)
  - 1133 species, 5214213 proteins, 143458 COGs
- Data extract: (Angela Wilkins and Daniel Konecki)
  - 7 species: human, mouse, zebrafish, *D. Melanogaster, C. Elegans*, yeast, *E. coli*
  - Only “experimentally confirmed” interactions
  - 59010 proteins represented

Protein-COG networks



Goal: identify missing or faulty protein-COG links
Limitation: Checking COGs by sequence works but is very slow

Red: query protein Cyan: proteins Blue: COGs

Michael Robinson

Using COG labels

- If two proteins are in the same COG, then they tend to be in **other** COGs **together** also

Start Protein	Species	D2	A2	GB	COGS in PPI-COG Network
ASIP	human	4	6.2	36	'COG0515', 'COG5023', 'COG5040', 'KOG0290', 'KOG0657', 'KOG0695', 'KOG0841', 'KOG1375', 'KOG1388', 'KOG1574', 'KOG3606', 'KOG3656', '**KOG4475**', 'KOG4643'
10090.ENSMUSP00000105319	mouse	6	4.2	21	'COG0515', 'COG5023', 'COG5040', 'KOG0290', 'KOG0657', 'KOG0695', 'KOG0841', 'KOG1375', 'KOG1388', 'KOG1574', 'KOG3606', 'KOG3656', '**KOG4222**', 'KOG4643'

- This holds for their neighbors as well

Only two differences

Key insight

- If two proteins have
  - similar interaction structure with neighboring proteins and
  - their neighbors are in similar COGs

  Then they probably are in the same COG

Key insight

- If two proteins have
  - similar interaction structure with neighboring proteins and
  - their neighbors are in similar COGs

  Then they probably are in the same COG

Base space

Goal: Narrow the search space of possible orthologs
Tool: Local topological and geometric invariants

Sheaf

Goal: "Zero in" on groups of proteins whose sequences are related, not to each other, but across species
Tool: *Consistency radius* of a sheaf of pseudometric spaces

What's new about this idea?

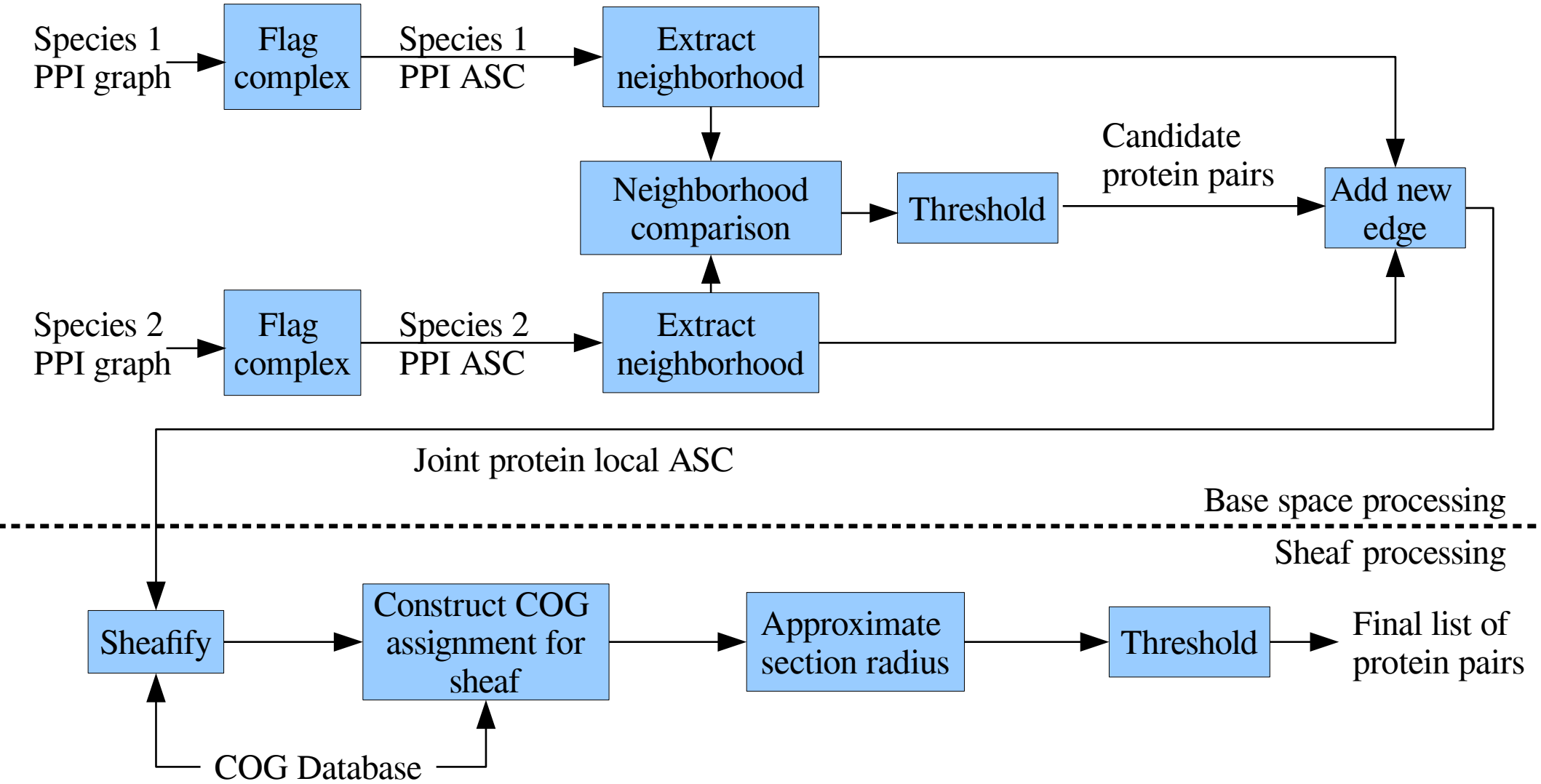
Usual procedure:

- Input:
  - Sequence data
  - Partial protein interactions
  - No COG information
- Output:
  - COG network

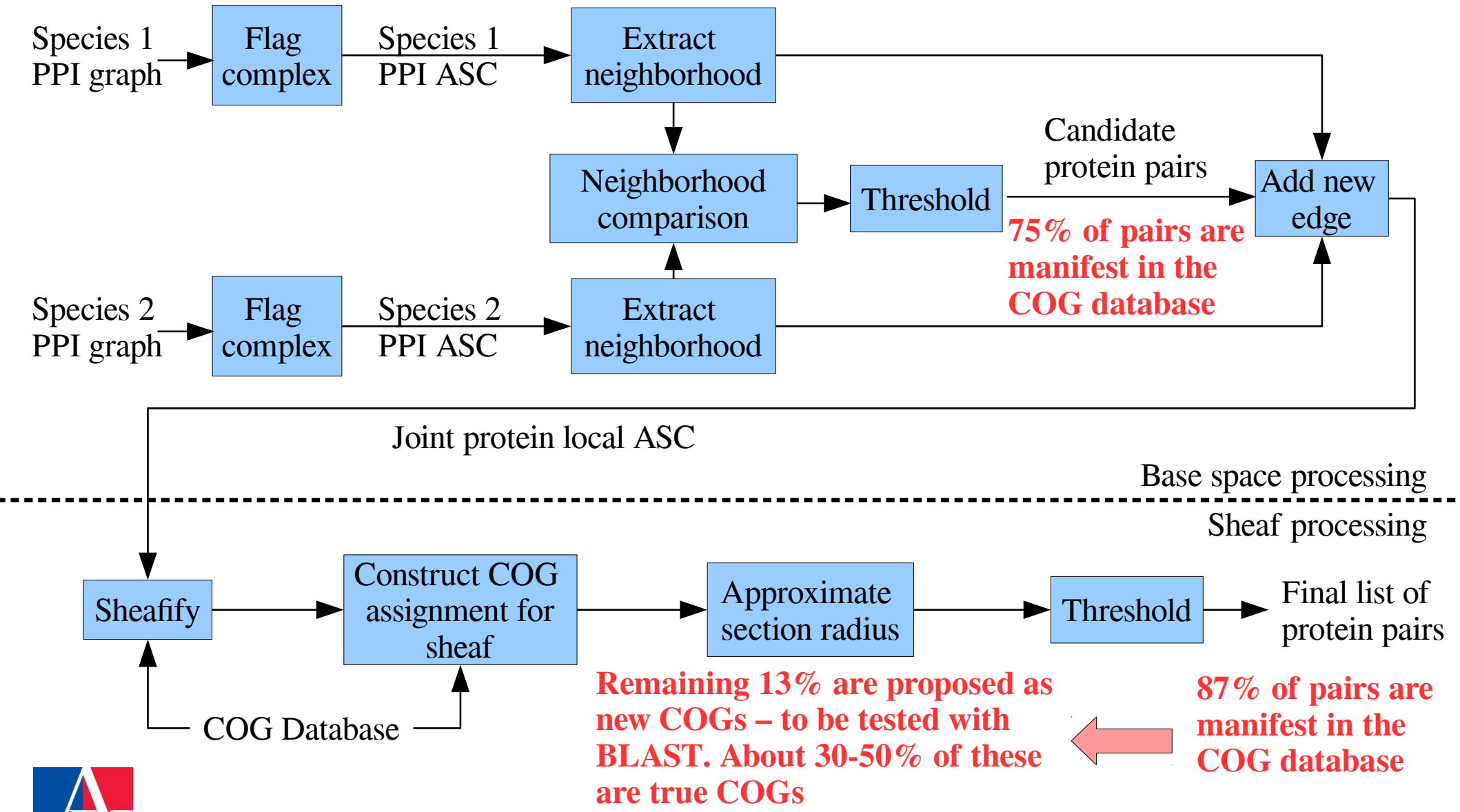
Our procedure:

- Input:
  - Protein interactions
  - Partial COG network
  - No sequences
- Output:
  - COG network

Process flowchart

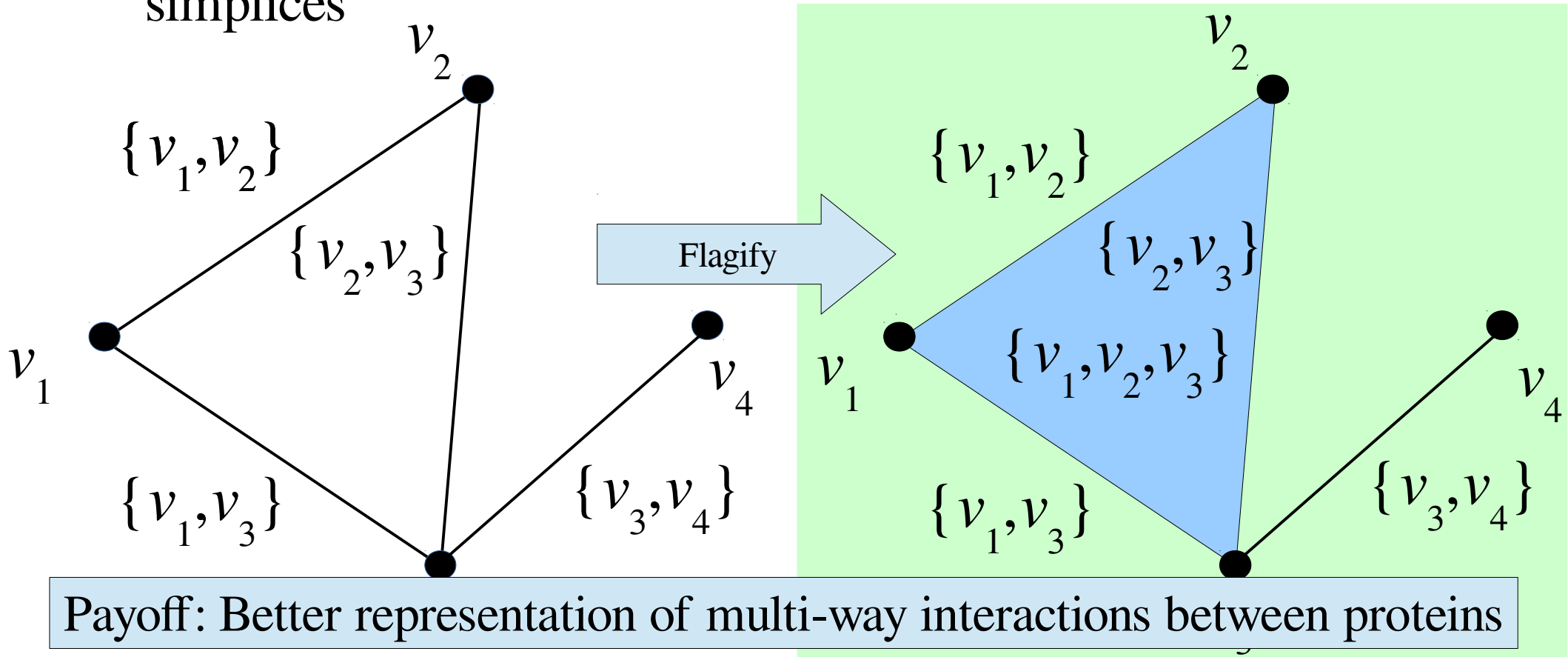


Process flowchart



Flag complex of PPI graph

- Vertices = proteins, Edges = interactions
- All *cliques* – an edge between every pair of vertices – become simplices



Payoff: Better representation of multi-way interactions between proteins

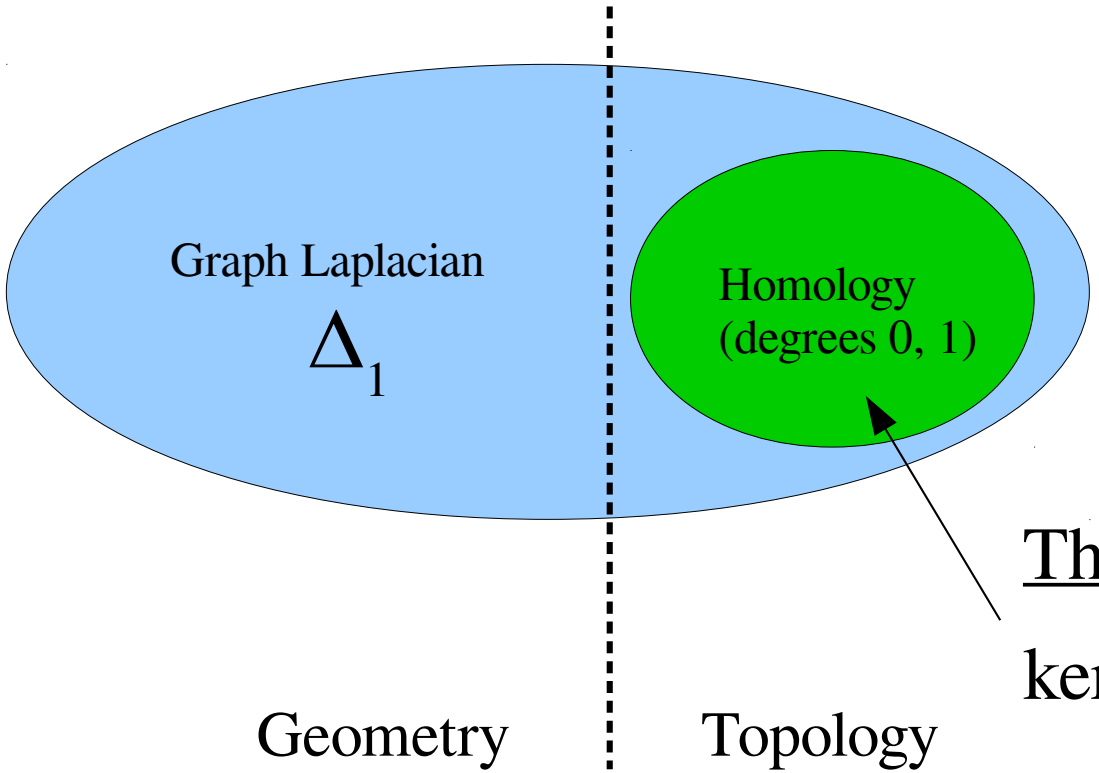
Matching metrics

- We look for pairs of proteins: one from each species with similar 2-hop neighborhoods
- There are several metrics available:

Graph Metric	Description
Vertex degree histogram	A list of vertex degree frequencies
Adjacency spectrum	Eigenvalues of graph adjacency matrix
Graph Laplacian spectrum	Eigenvalues of the Laplacian matrix where a Laplacian matrix is the adjacency matrix subtracted from the diagonal matrix of vertex degrees
Graph density (undirected graph)	Density = $(2m) / (n(n-1))$, where n = # edges, m = # vertices
Graph Betti number (connected graph)	Graph Betti = $n - m + 1$, where n = # edges, m = # vertices

Aside: Homology and spectra

- In a graph, the graph Laplacian Δ_1 determines homology, so it's convenient and widely used

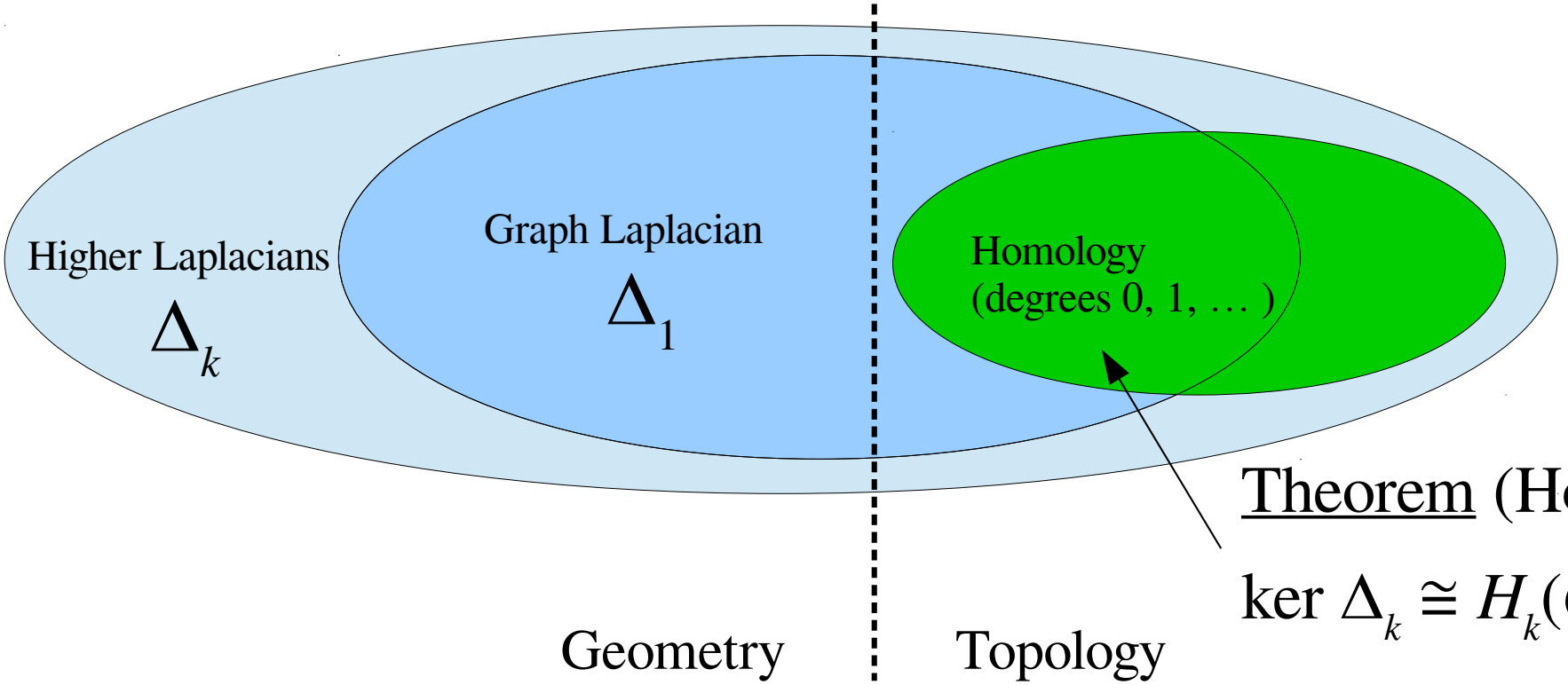


Theorem (Hodge):

$\ker \Delta_k \cong H_k(C_\bullet, \partial_\bullet)$

Aside: Homology and spectra

- For cell complexes, the graph Laplacian and homology are different, but related
- There are "higher" Laplacians that determine homology, but they aren't much used* in data science



Theorem (Hodge):

$\ker \Delta_k \cong H_k(C_\cdot, \partial_\cdot)$

* I'm not sure why, actually! But... we aren't either yet :-(

Refining the search

- How well are local network invariants from a COG's proteins correlated across species?

Graph Metric	Topological?	Pearson Correlation
Second bin degree histogram (D2)	Yes	0.9046
Second adjacency eigenvalue (A2)	Partially	0.8823
Second Laplacian eigenvalue (L2)	Partially	0.3596
Graph density (GD)	No	0.5634
Graph Betti number (GB)	Yes	0.8840

Local topology is a strong indicator, but is not conclusive...
Remember we're looking at 50000+ proteins!

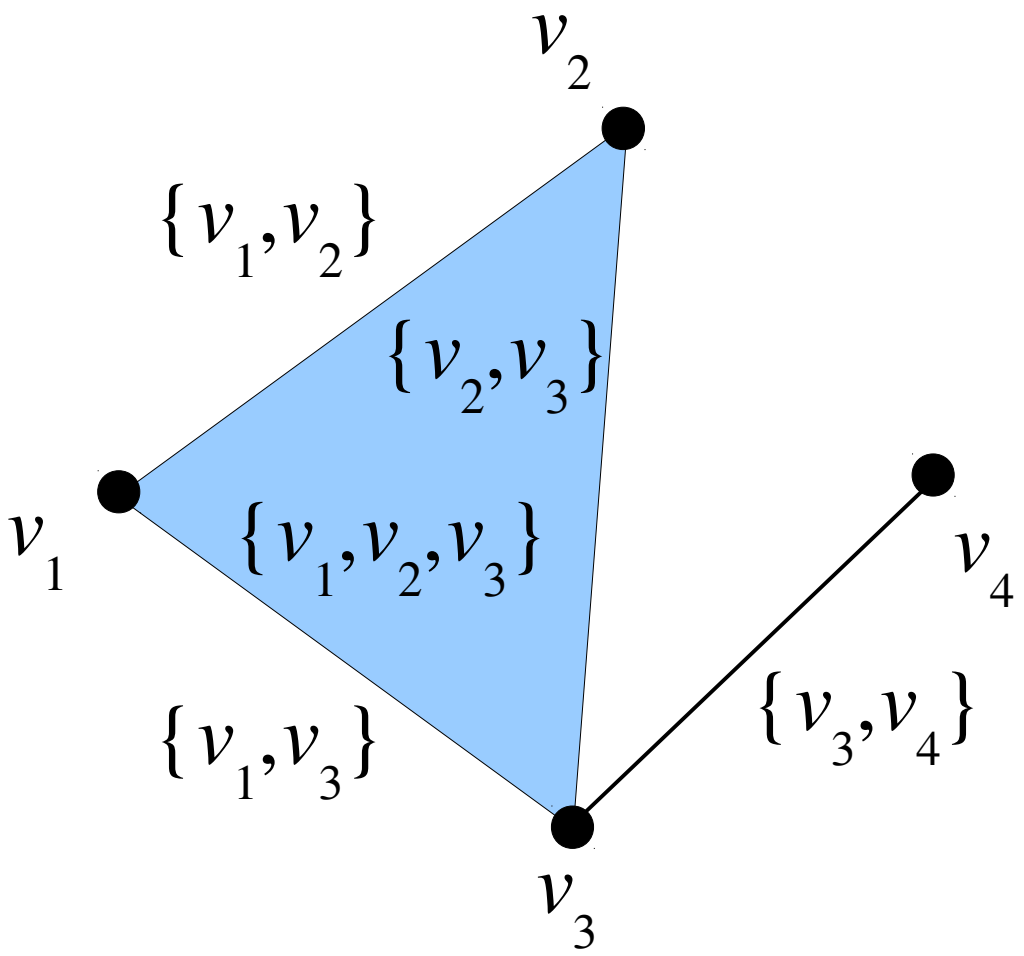
- The local topology and geometry of the protein-COG network greatly reduces the search space

Local sections

- The mantra of algebraic topology is “local to global”
  - Poor scaling (usually cubic in the number of simplices)
  - Requires linear algebra (usually good, but not always)
  - Real data usually can’t be globalized due to errors
- Very little effort has been expended by others about “partially global” results: *local sections* of sheaves
- We have recently been looking at local sections
  - Discovery: Interesting combinatorics is present!
  - Payoff: Partially global results are more realistic, and easier to compute

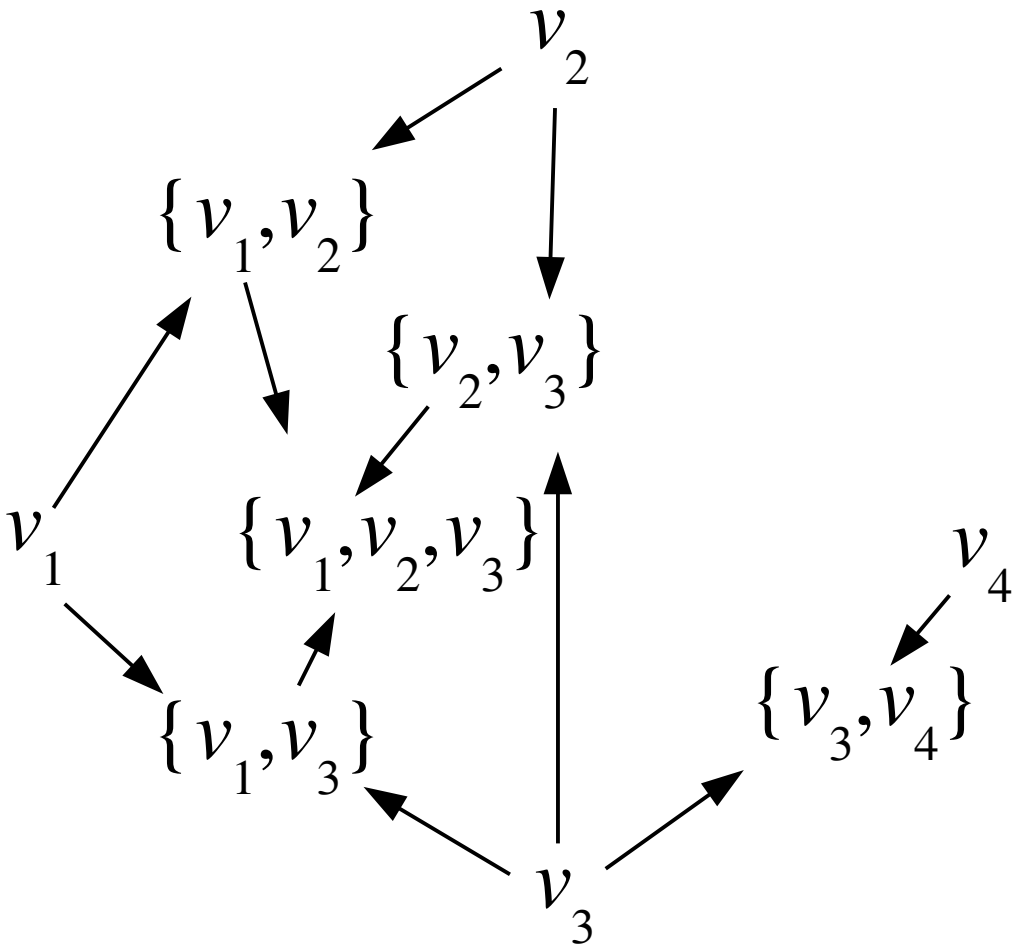
Simplicial complexes

- An *abstract simplicial complex* consists of *simplices* (tuples of vertices)



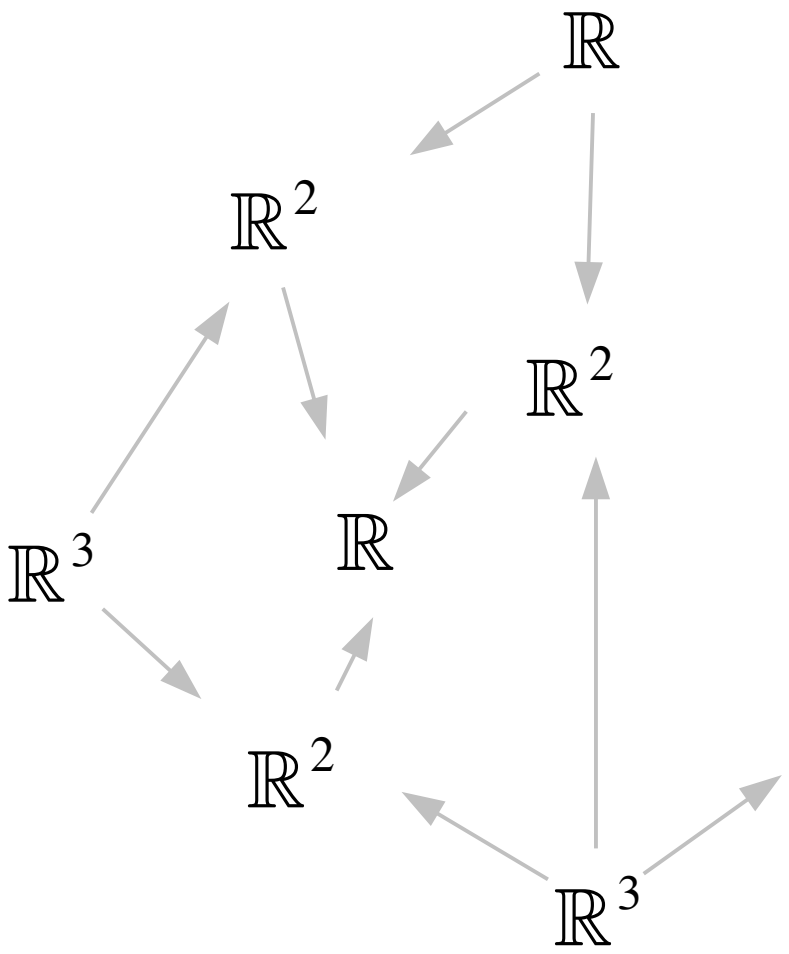
Simplicial complexes

- The *attachment diagram* shows how simplices fit together

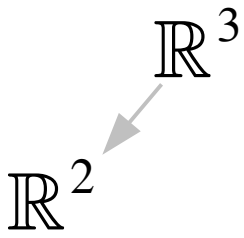


A *sheaf* is ...

- A set assigned to each simplex and ...



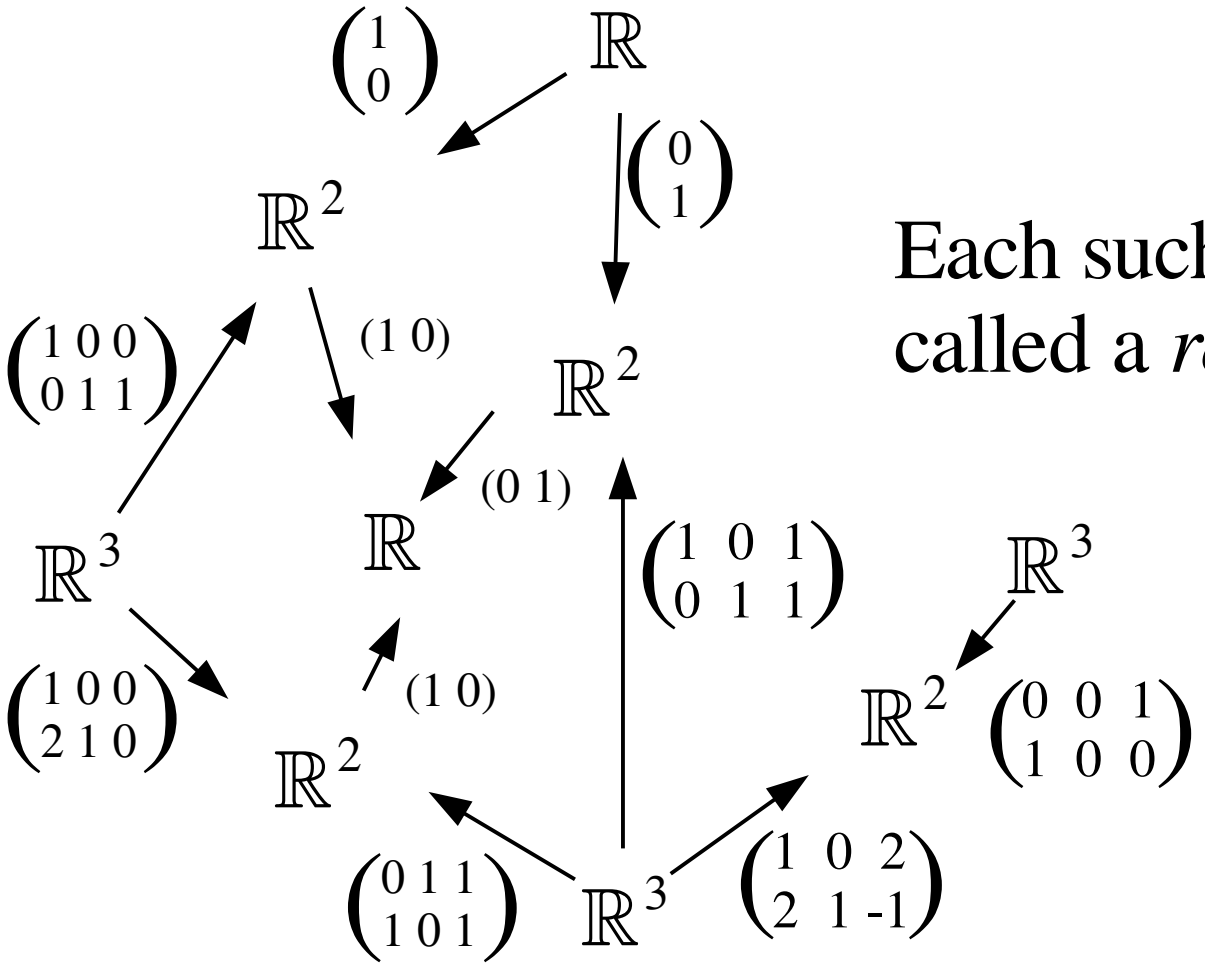
Each such set is called the *stalk* over its simplex



This is a sheaf **of** vector spaces **on** a simplicial complex

A *sheaf* is ...

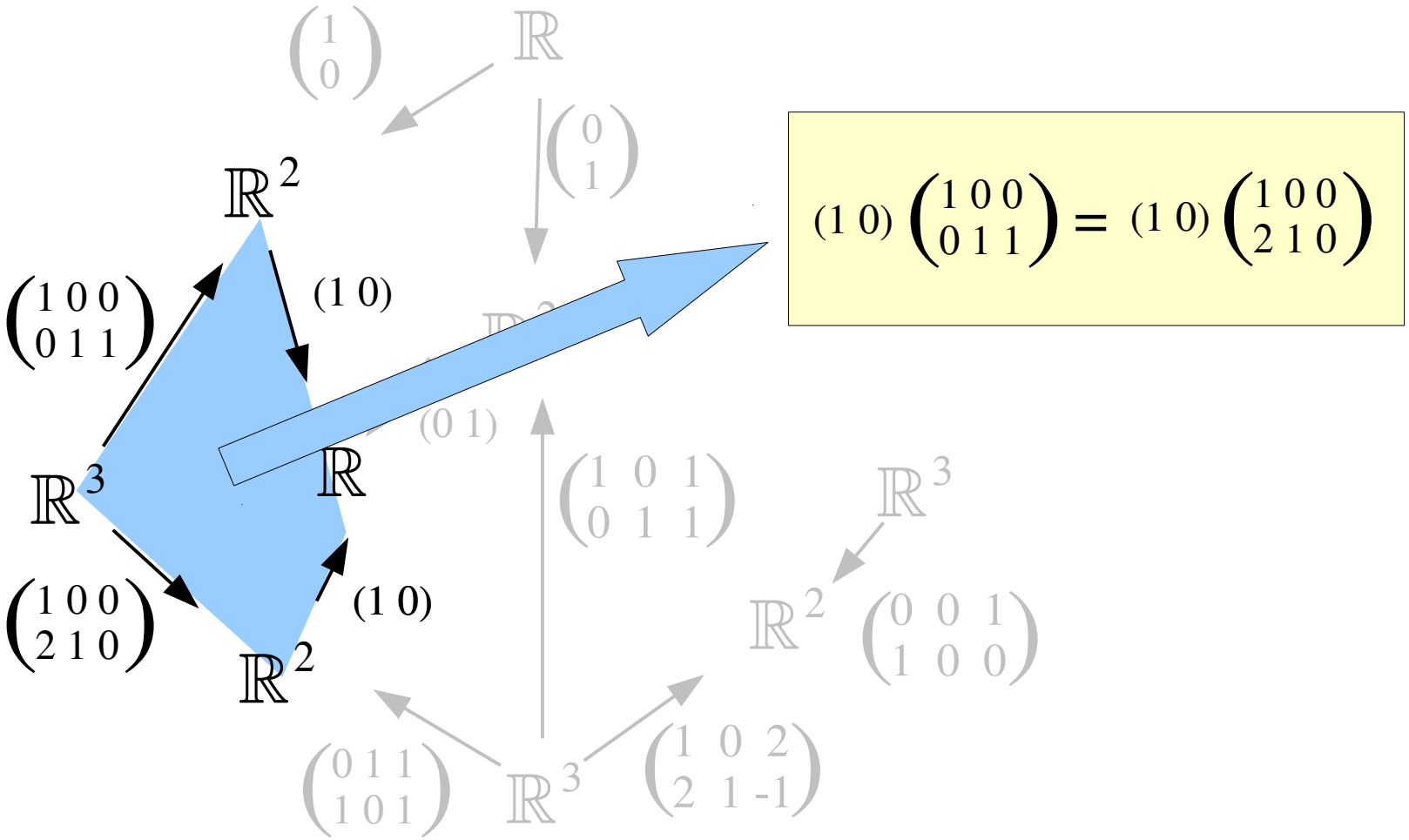
- … a function assigned to each simplex inclusion



Each such function is called a *restriction*

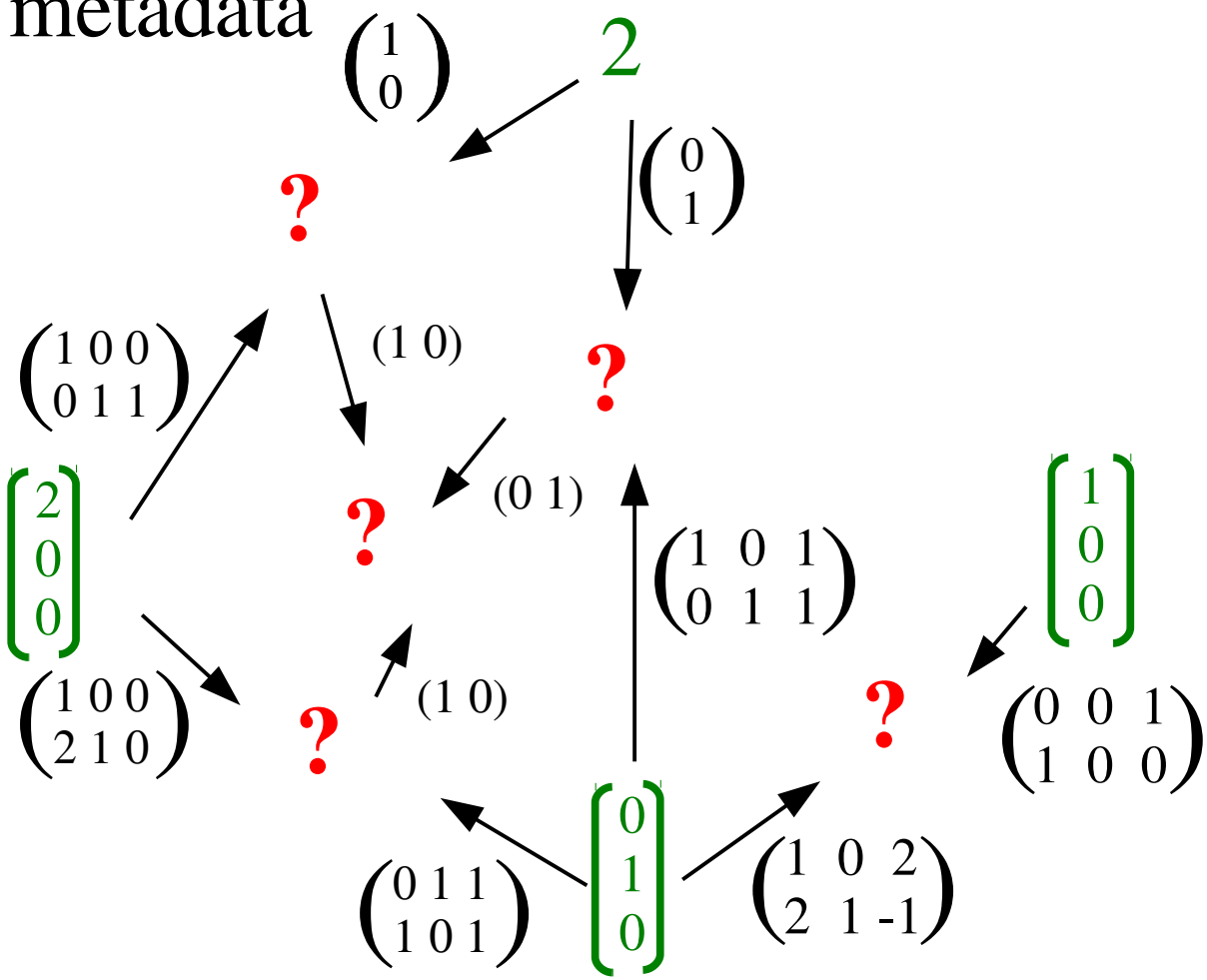
A *sheaf* is ...

- … so the diagram commutes.



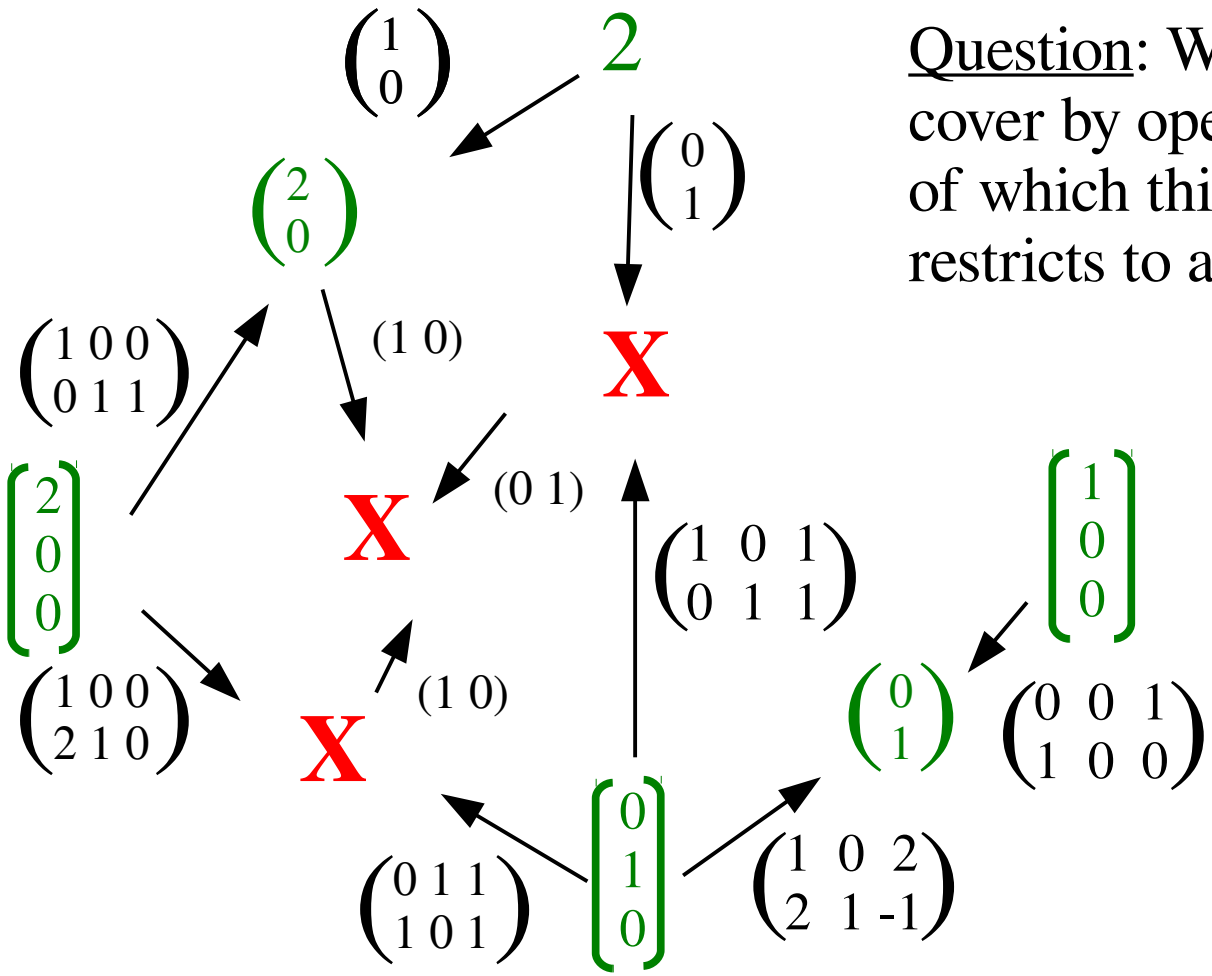
Consider a vertex assignment

- Values are placed at vertices only, corresponding to protein metadata



Consider a vertex assignment

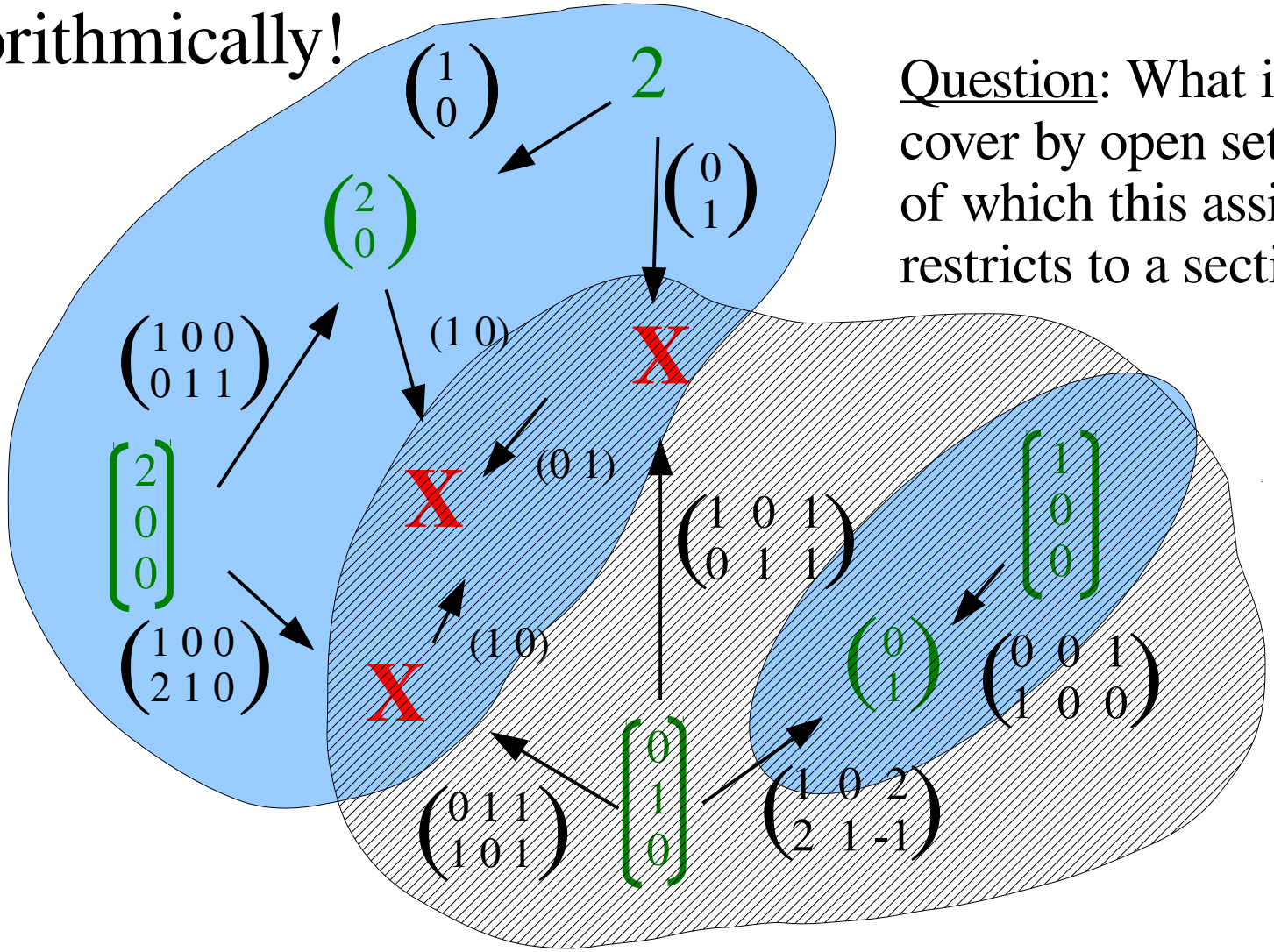
- In some places there is consistency, but not all



Question: What is the best cover by open sets, on each of which this assignment restricts to a section?

Maximal covers of local sections

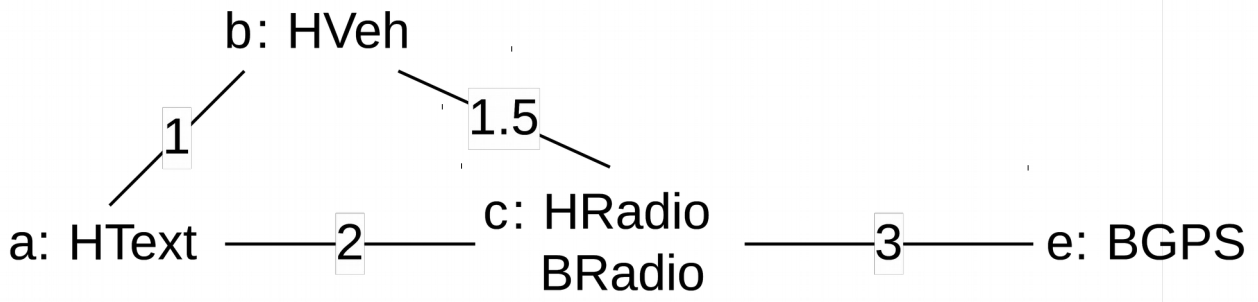
- Theorem: (Praggastis) we can compute the cover algorithmically!



Question: What is the best cover by open sets, on each of which this assignment restricts to a section?

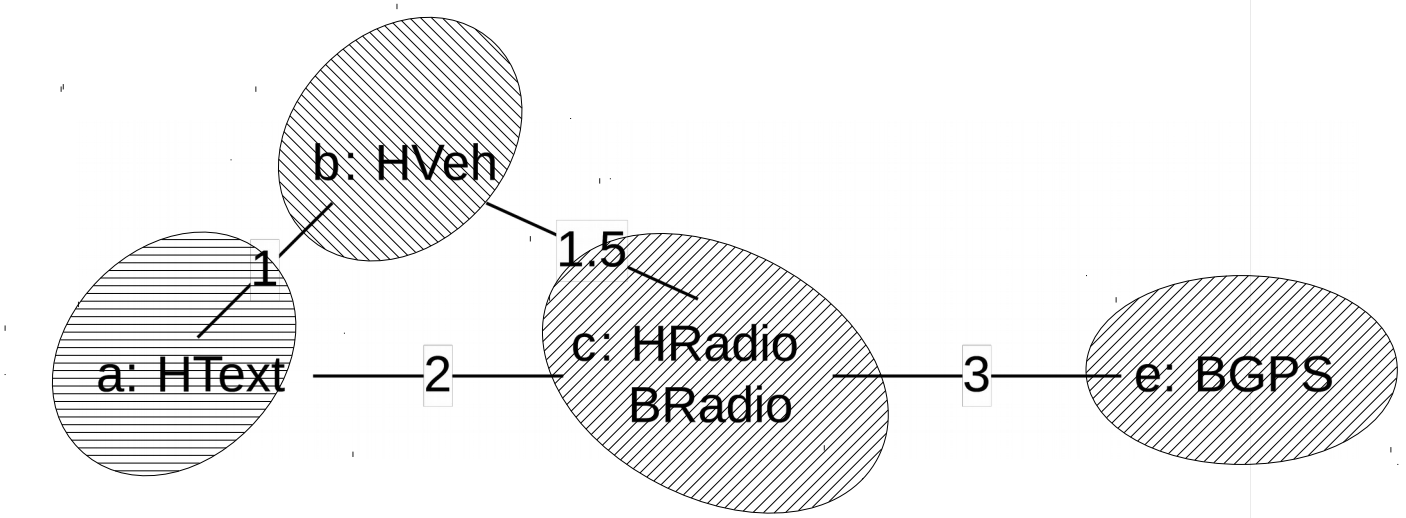
Sectional Analysis

- Set of observations: $d(a,b)=1$, $d(b,c)=1.5$, $d(a,c)=2$, $d(c,e)=3$
- Max error (a radius): $\varepsilon^* = \max(d(a,b),d(b,c),d(a,c),d(c,e))/2 = 1.5$
- Sequence of radii: (0.5,0.75,1.0,1.5)
- Sectional filtration on ε



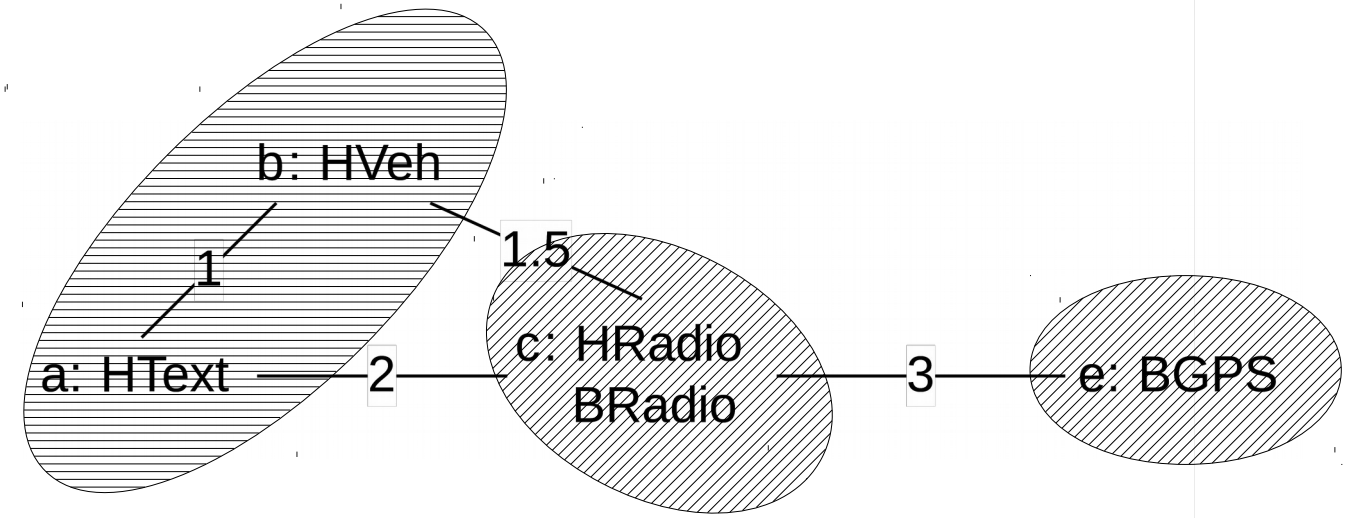
Sectional Analysis

- Set of observations: $d(a,b)=1$, $d(b,c)=1.5$, $d(a,c)=2$, $d(c,e)=3$
- Max error (a radius): $\varepsilon^* = \max(d(a,b),d(b,c),d(a,c),d(c,e))/2 = 1.5$
- Sequence of radii: (0.5,0.75,1.0,1.5)
- Sectional filtration on ε
  - 0.0: a/b/c/e



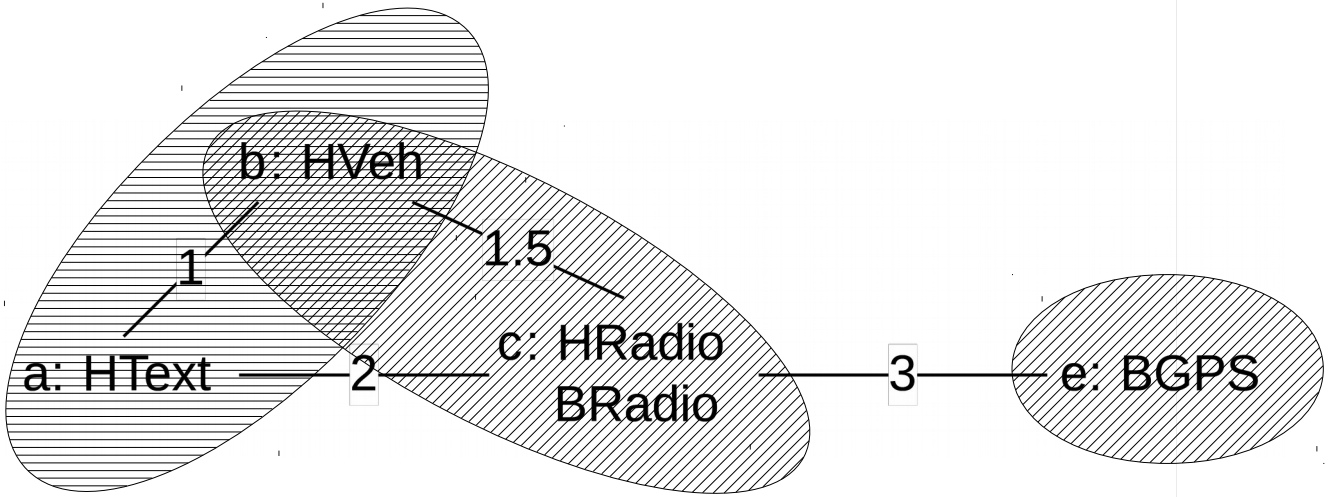
Sectional Analysis

- Set of observations: $d(a,b)=1$, $d(b,c)=1.5$, $d(a,c)=2$, $d(c,e)=3$
- Max error (a radius): $\varepsilon^* = \max(d(a,b),d(b,c),d(a,c),d(c,e))/2 = 1.5$
- Sequence of radii: (0.5,0.75,1.0,1.5)
- Sectional filtration on ε
  - 0.0: a/b/c/e
  - 0.5: ab/c/e



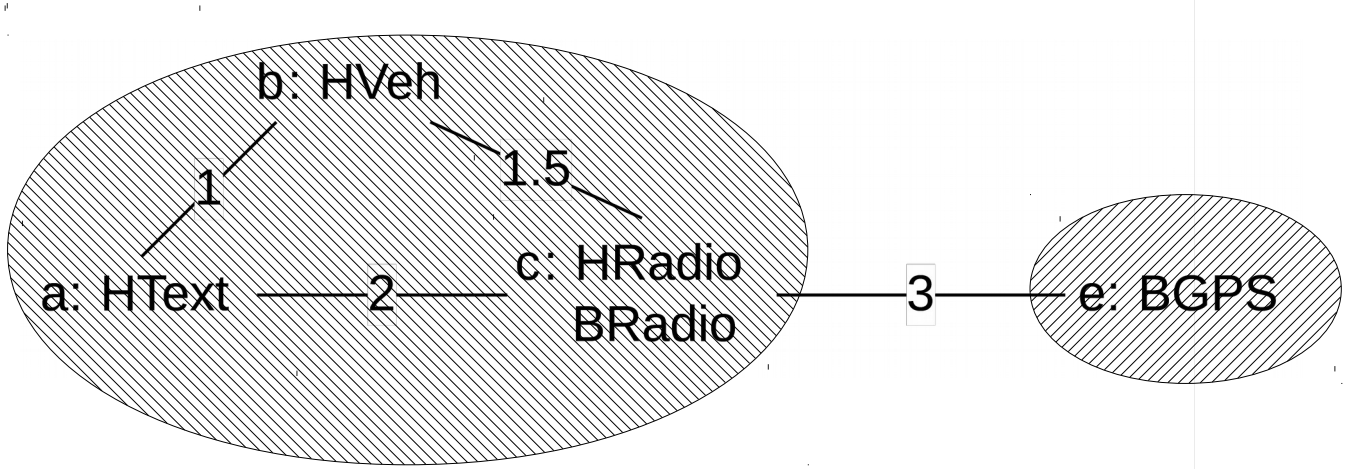
Sectional Analysis

- Set of observations: $d(a,b)=1$, $d(b,c)=1.5$, $d(a,c)=2$, $d(c,e)=3$
- Max error (a radius): $\varepsilon^* = \max(d(a,b),d(b,c),d(a,c),d(c,e))/2 = 1.5$
- Sequence of radii: (0.5,0.75,1.0,1.5)
- Sectional filtration on ε
  - 0.0: a/b/c/e
  - 0.5: ab/c/e
  - 0.75: ab/bc/e



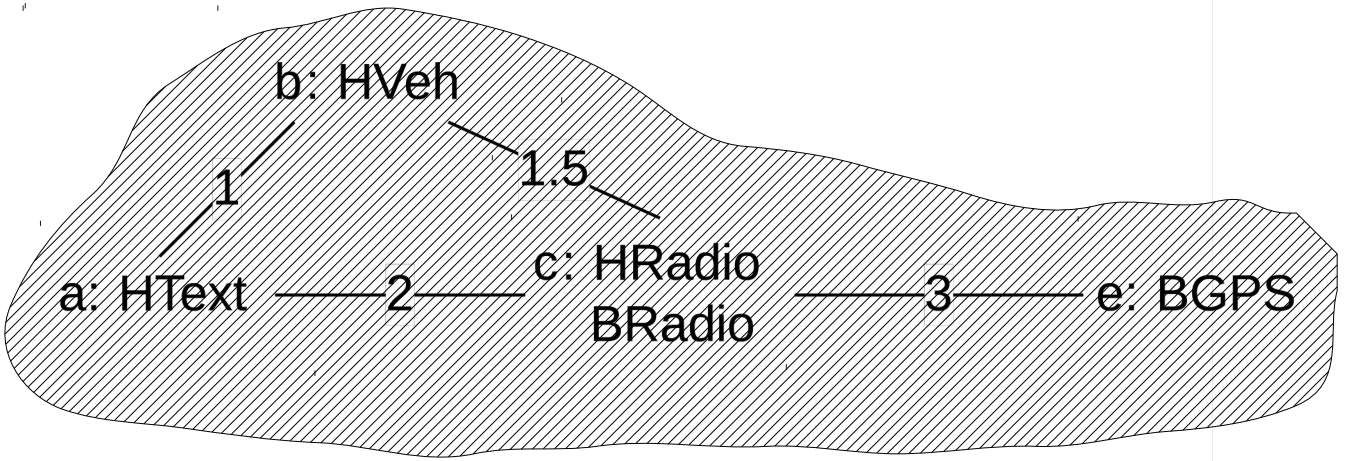
Sectional Analysis

- Set of observations: $d(a,b)=1$, $d(b,c)=1.5$, $d(a,c)=2$, $d(c,e)=3$
- Max error (a radius): $\varepsilon^* = \max(d(a,b),d(b,c),d(a,c),d(c,e))/2 = 1.5$
- Sequence of radii: (0.5,0.75,1.0,1.5)
- Sectional filtration on ε
  - 0.0: a/b/c/e
  - 0.5: ab/c/e
  - 0.75: ab/bc/e
  - 1.0: abc/e



Sectional Analysis

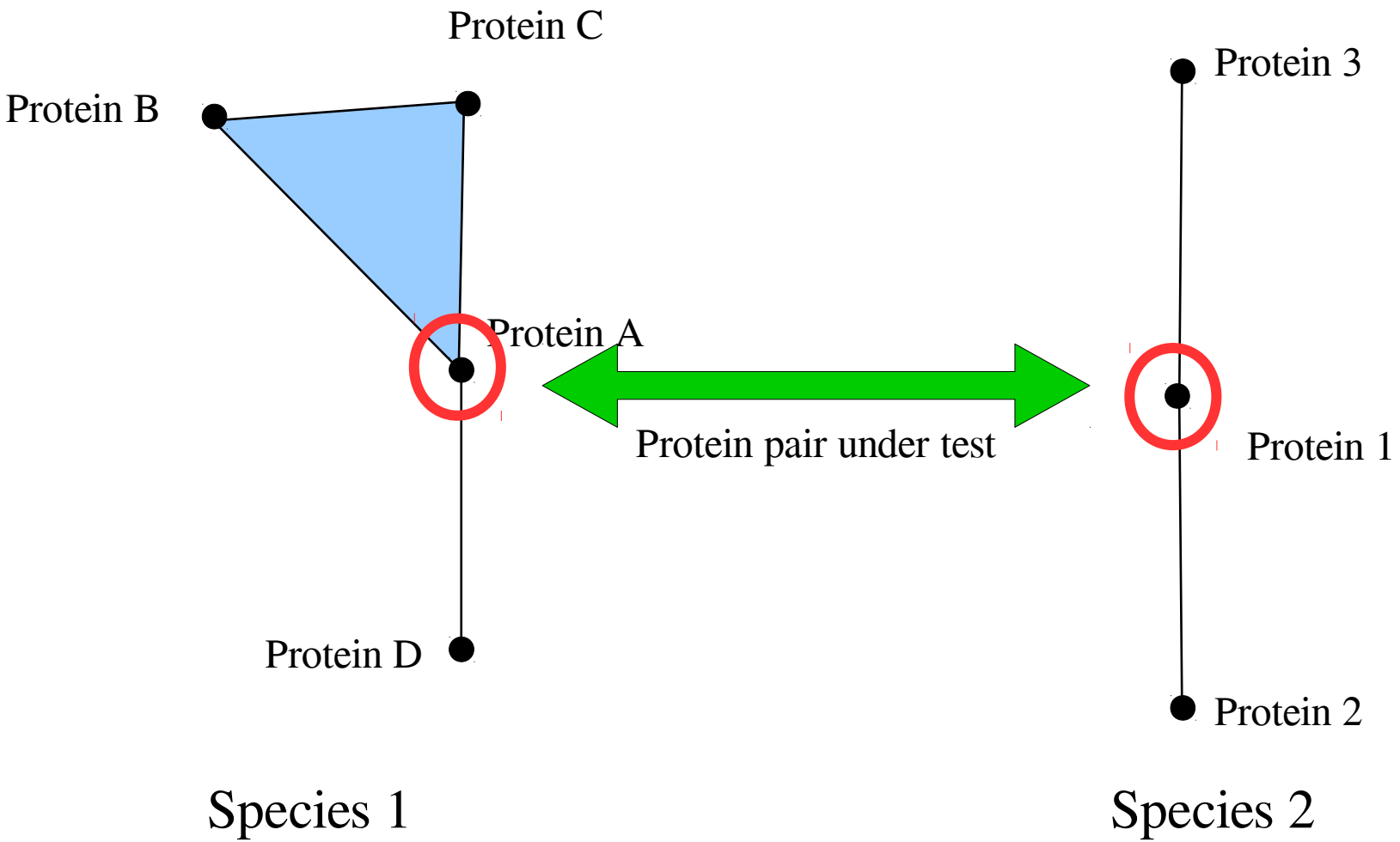
- Set of observations: $d(a,b)=1$, $d(b,c)=1.5$, $d(a,c)=2$, $d(c,e)=3$
- Max error (a radius): $\varepsilon^* = \max(d(a,b),d(b,c),d(a,c),d(c,e))/2 = 1.5$
- Sequence of radii: (0.5,0.75,1.0,1.5)
- Sectional filtration on ε
  - 0.0: a/b/c/e
  - 0.5: ab/c/e
  - 0.75: ab/bc/e
  - 1.0: abc/e
  - 1.5: abce



The *consistency radius* is the smallest threshold yielding global consistency
Theorem: (Nowak) This can be computed algorithmically!

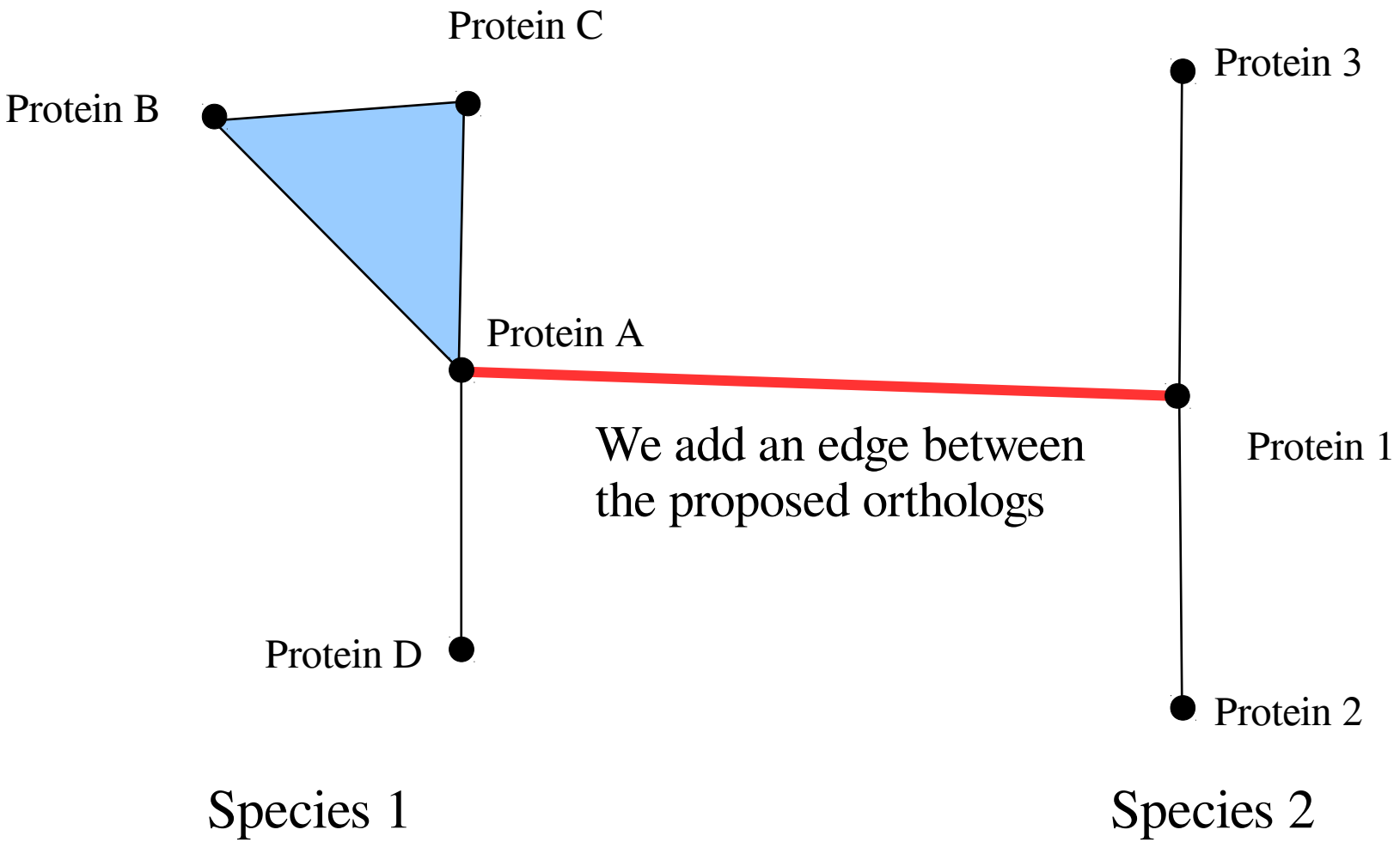
Local PPI complexes

NB: we use the 2-hop neighborhood, even though I'm only showing the 1-hop neighborhood



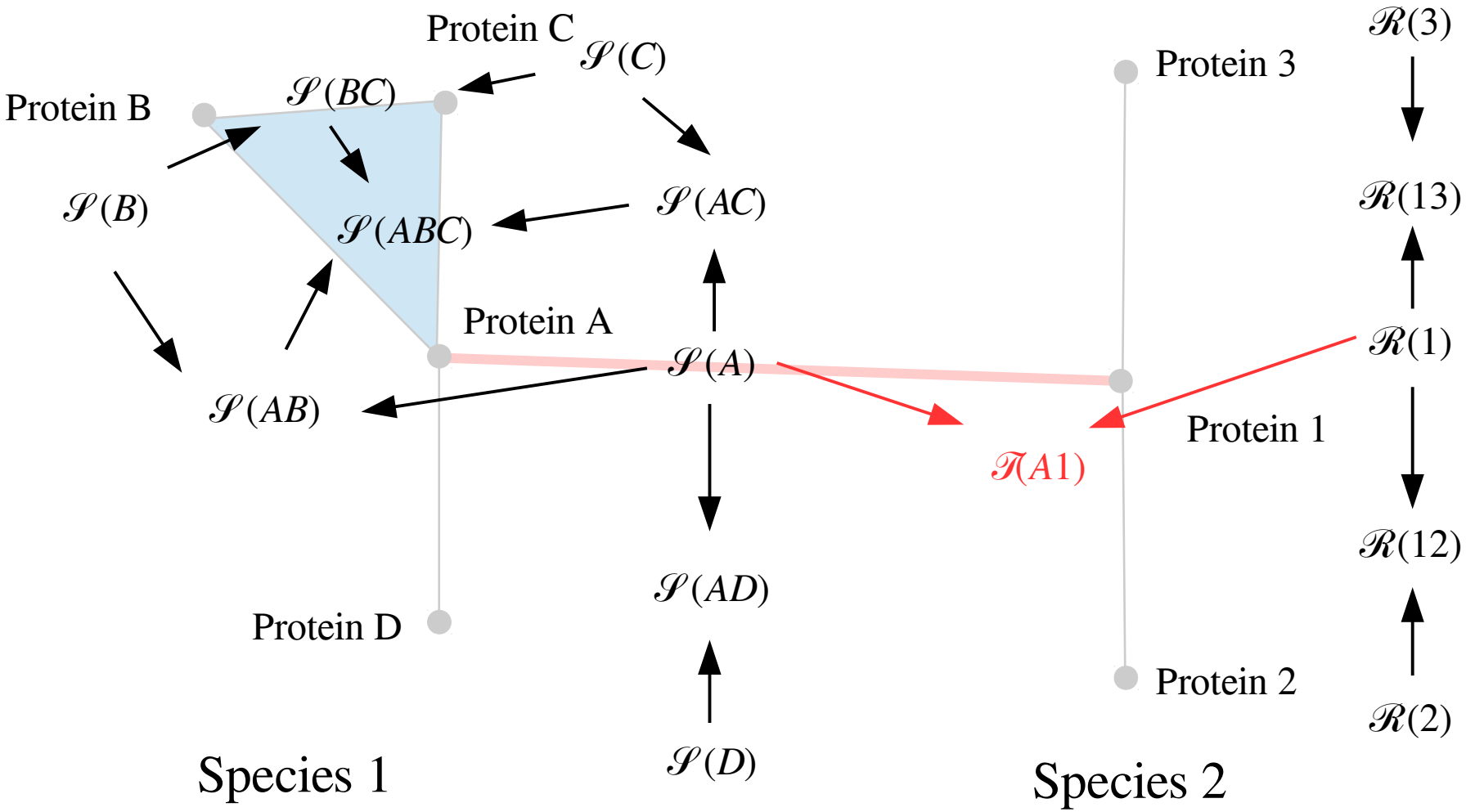
Joint local PPI complex

NB: we use the 2-hop neighborhood, even though I'm only showing the 1-hop neighborhood



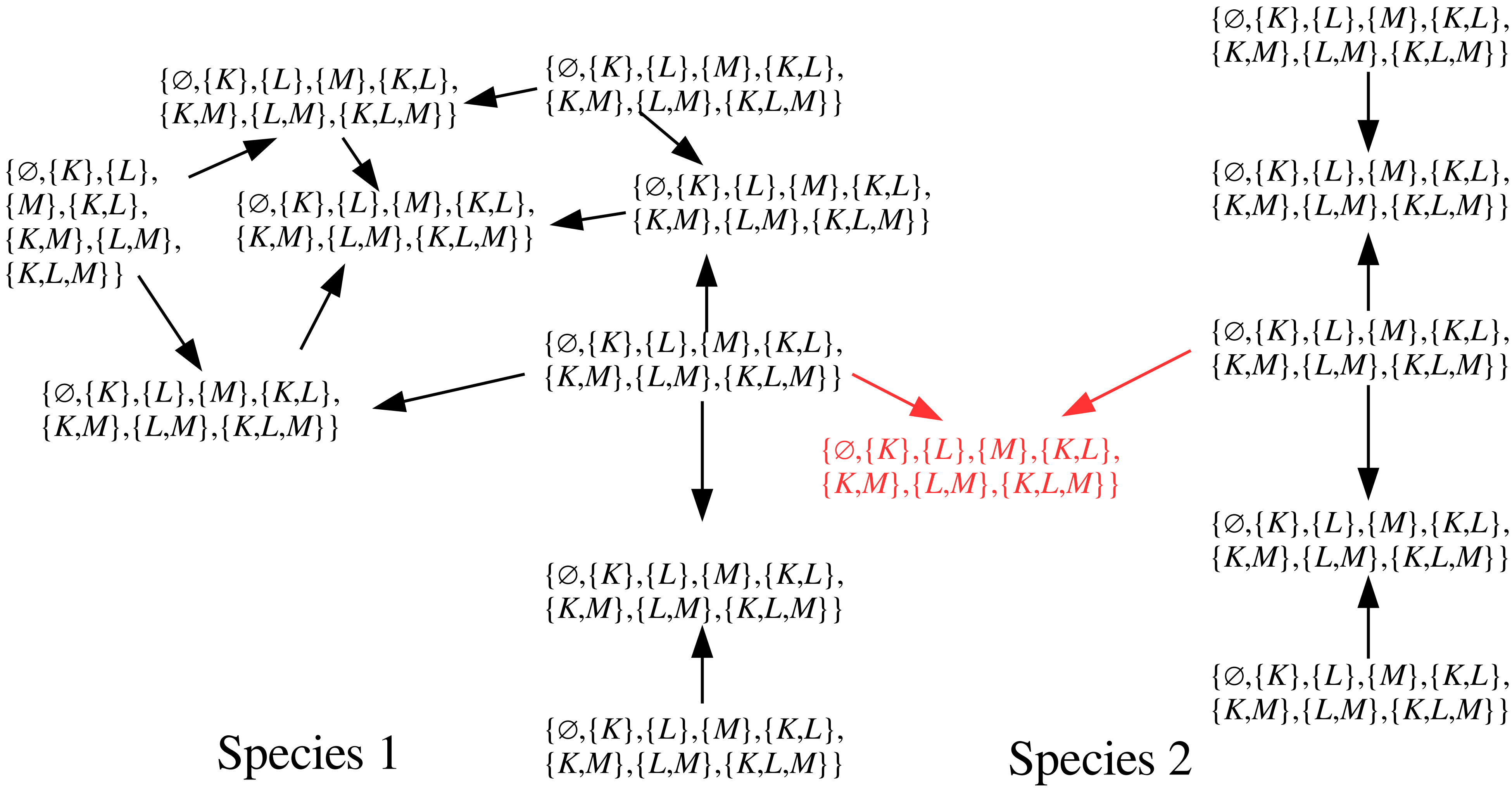
Sheaf of COG label sets

NB: we use the 2-hop neighborhood, even though I'm only showing the 1-hop neighborhood



Sheaf of COG label sets

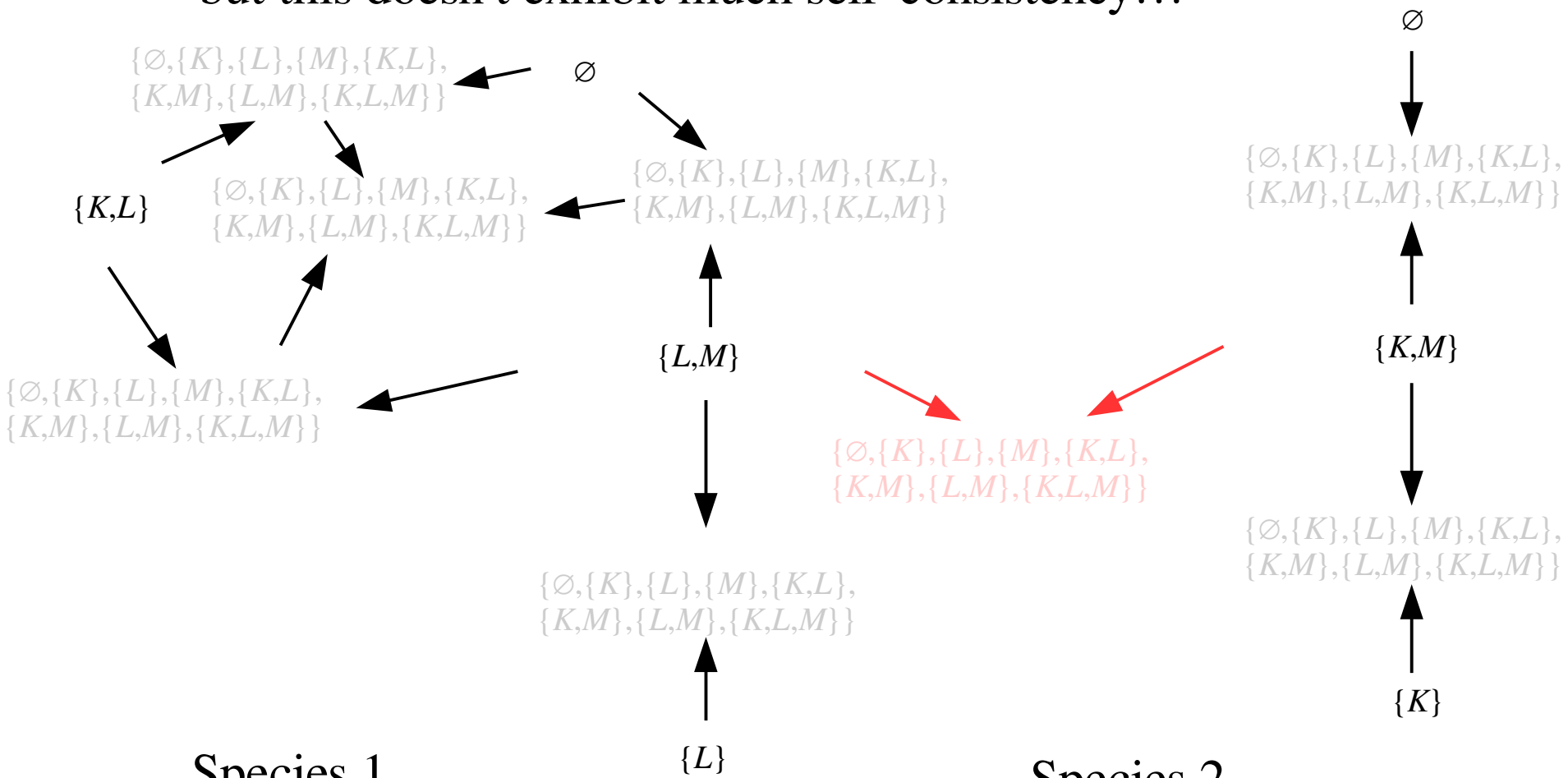
Three known COGs: K, L, M



All restrictions are identity functions...

Sheaf of COG label sets

The COG database consists of a vertex assignment, like so…
but this doesn't exhibit much self-consistency…



Species 1

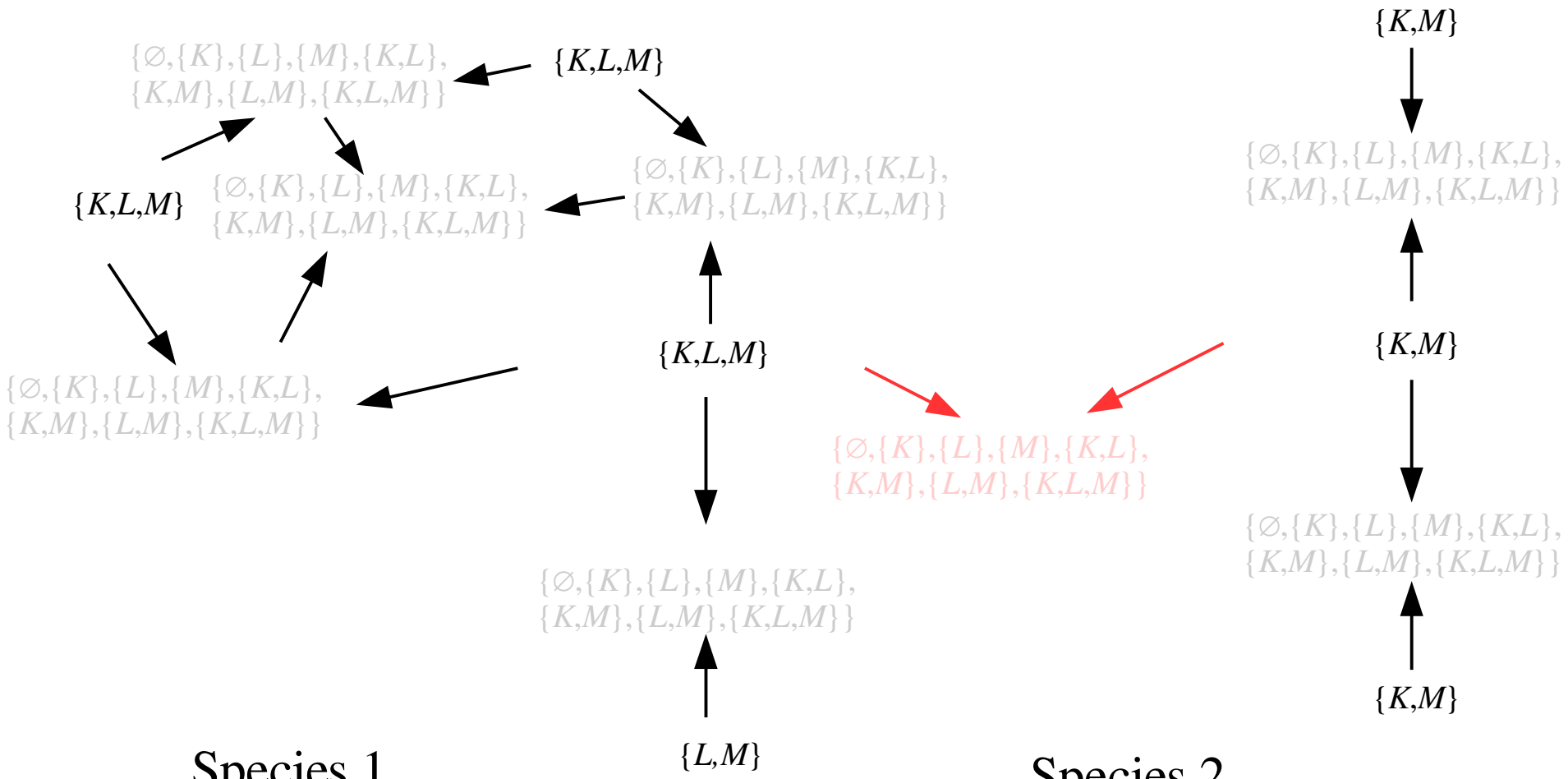
Species 2

All restrictions are identity functions...

Michael Robinson

Sheaf of COG label sets

… so instead assign the set of COGs of each protein and its neighbors…



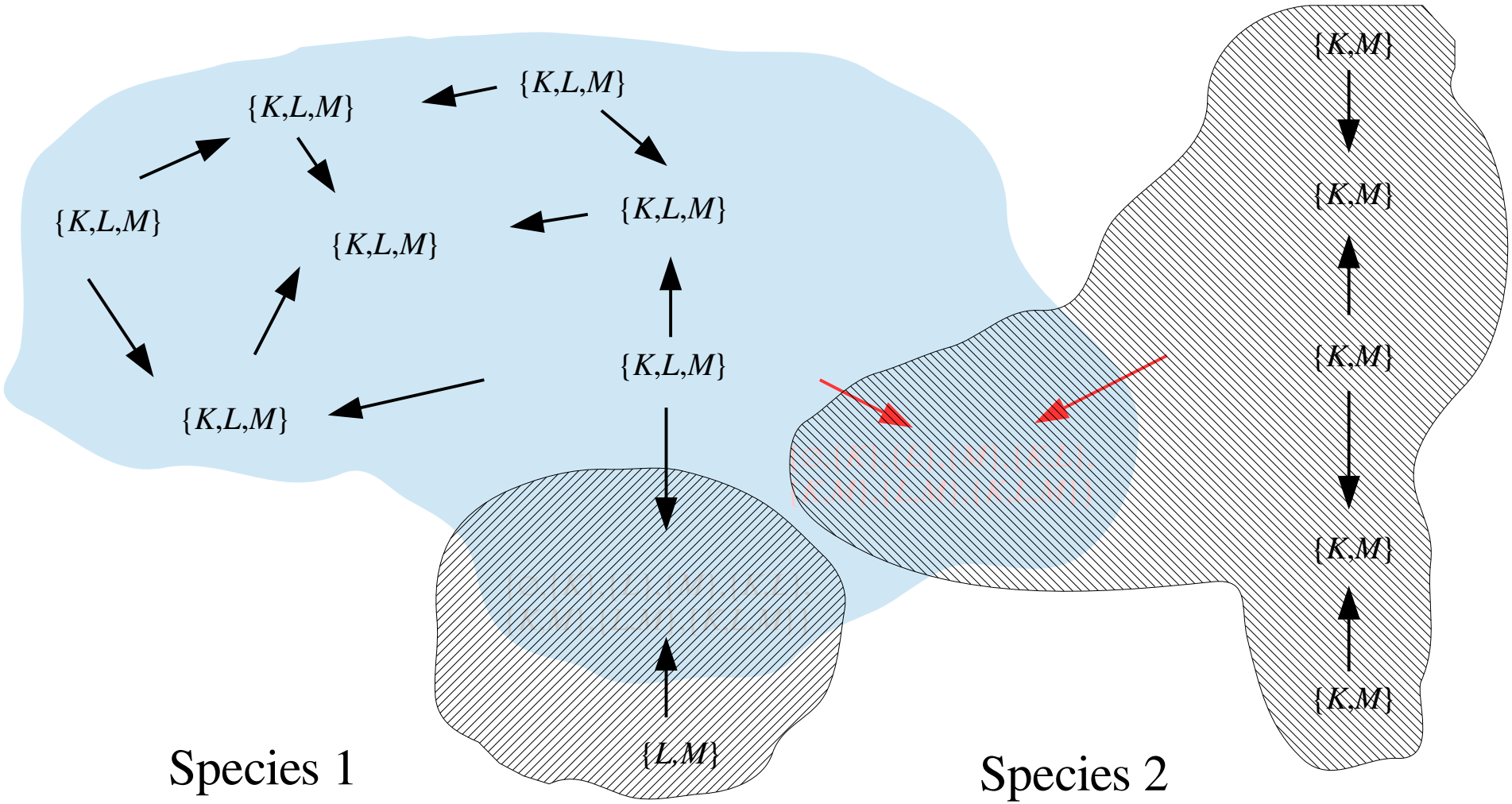
Species 1

Species 2

All restrictions are identity functions...

Sheaf of COG label sets

… Extend to maximal local sections. If not a global section...

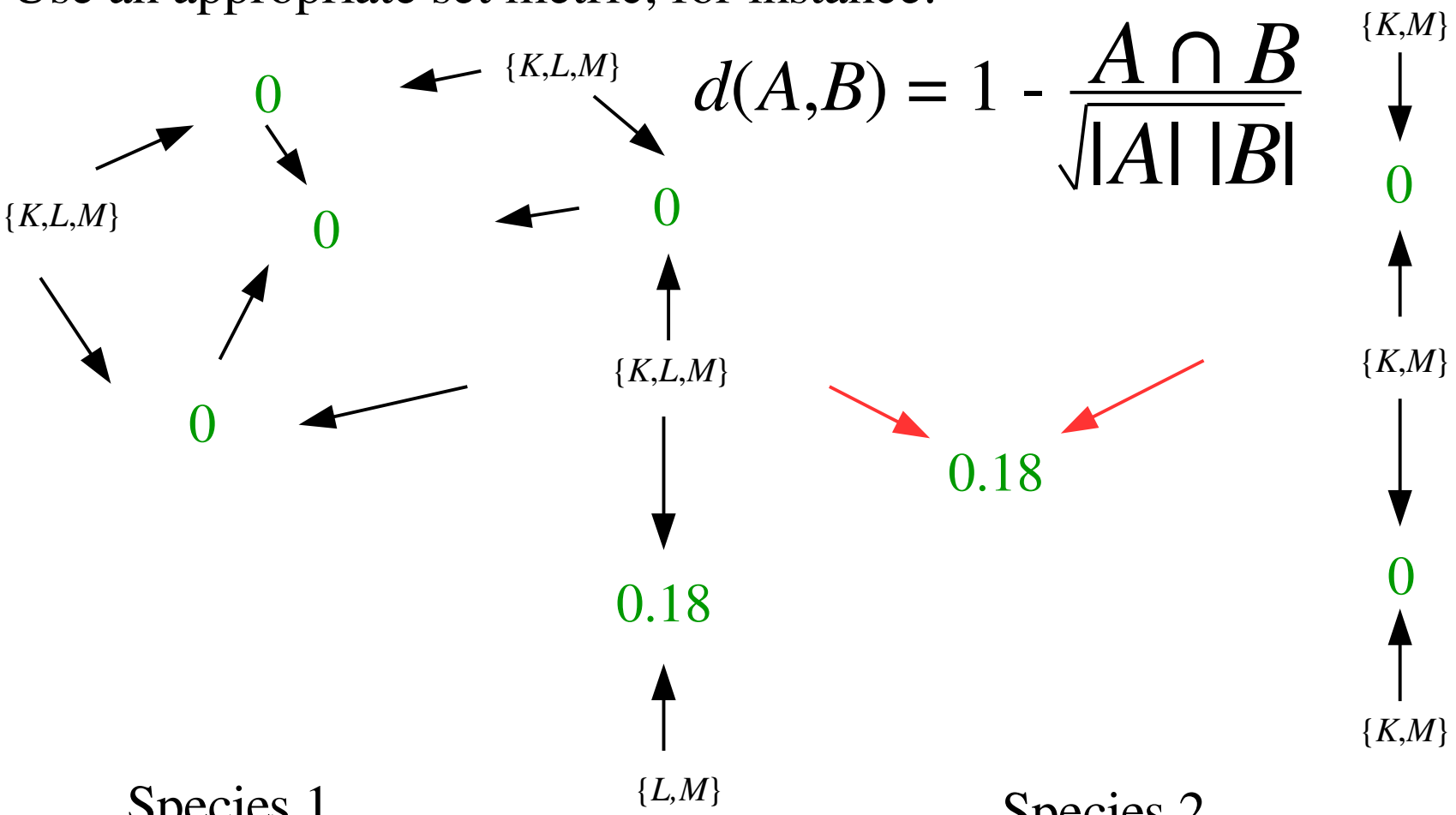


All restrictions are identity functions...

Sheaf of COG label sets

… compute the *consistency radius*
Use an appropriate set metric, for instance:

$$d(A,B) = 1 - \frac{A \cap B}{\sqrt{|A|\,|B|}}$$



Validation process

Sheaf-based ortholog algorithm

Prediction

Validation

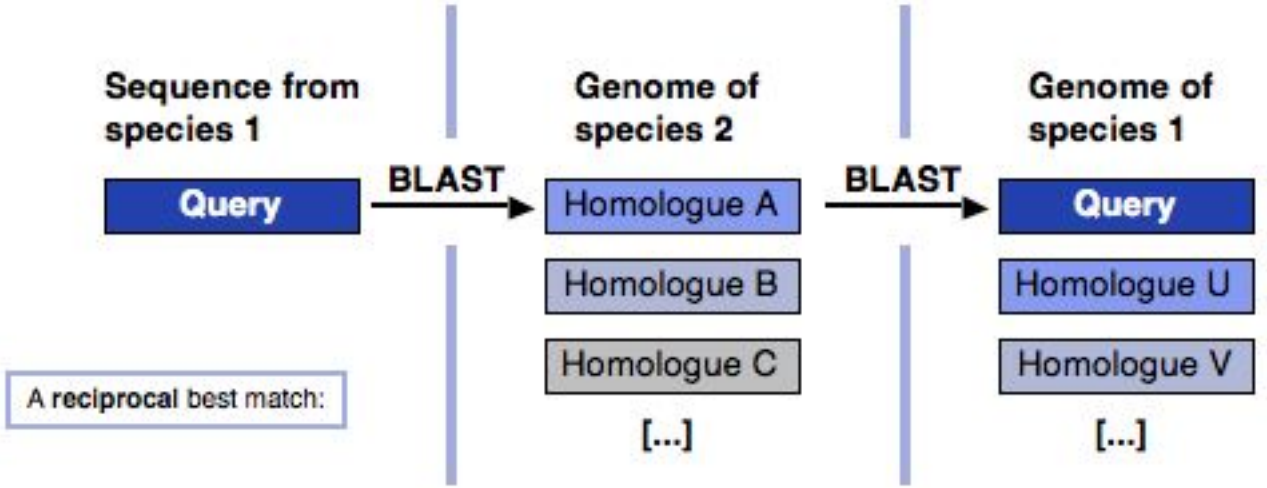
Look up sequences

Reciprocal BLAST

Threshold

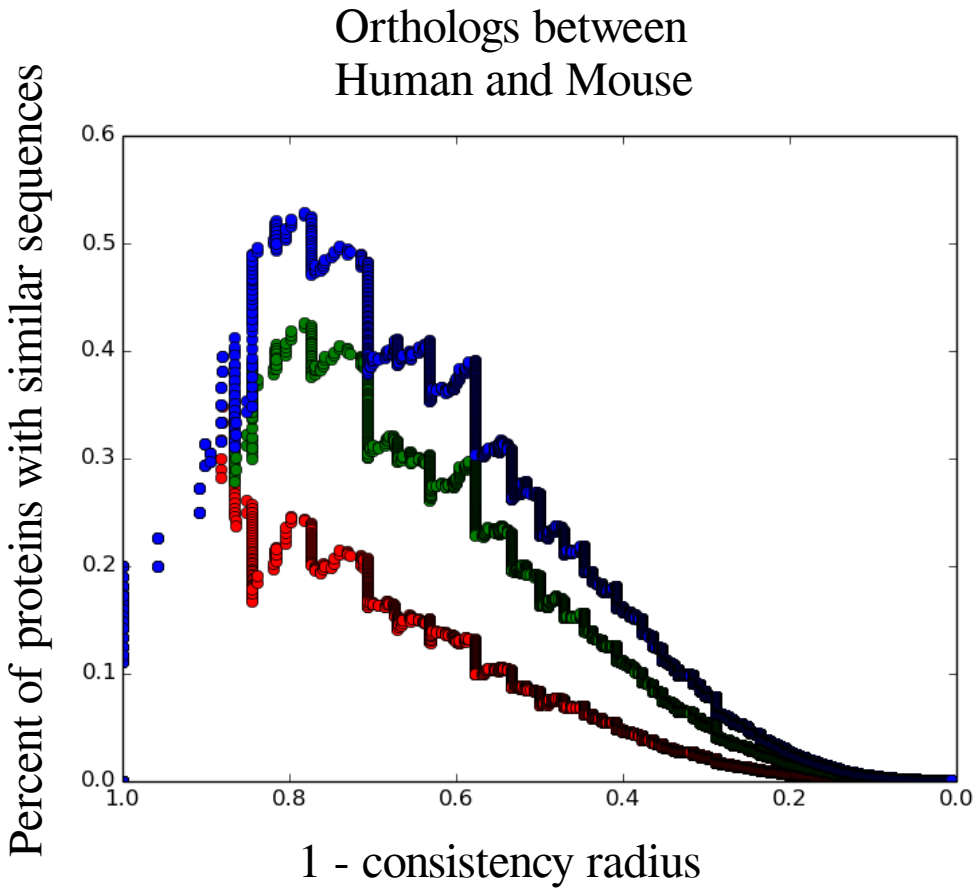
Match?

	A	B	C	D	E	F
1	COGName	Human Protein	Mouse Protein	Metric	Cog Set Length Human	Cog Set Leng
2	NOG40074	CLDN19	10090.ENSMUSP0000008133	1	2	
3	KOG3547	BEST1	10090.ENSMUSP0000011305	1	3	
4	NOG28428	C1orf116	10090.ENSMUSP0000004617	1	2	
5	NOG28592	CBLN3	10090.ENSMUSP0000007049	1	3	
6	NOG293748	CRYGC	10090.ENSMUSP0000008461	1	5	
7	NOG294965	[illegible]	10090.ENSMUSP0000003897	1	4	
8	NOG3897	CLDN16	[illegible]	1	2	
9	NOG392[illegible]	ARMC12	10090.ENSMUSP0000002506	1	2	
10	KOG3538	[illegible]	[illegible]	1	5	
11	COG1012	ALDH4A1	10090.ENSMUSP0000004382	1	10	
12	COG1131	ABCA13	10090.ENSMUSP0000004046	1	5	
13	KOG3500	ATP6V0E1	10090.ENSMUSP0000004411	1	8	
14	NOG40269	AKAP4	10090.ENSMUSP0000005096	1	7	
15	NOG40339	ART1	10090.ENSMUSP0000003330	1	4	
16	NOG40437	CCDC87	10090.ENSMUSP0000008602	1	7	
17	KOG1542	CTSW	10090.ENSMUSP0000010930	1	3	



Michael Robinson

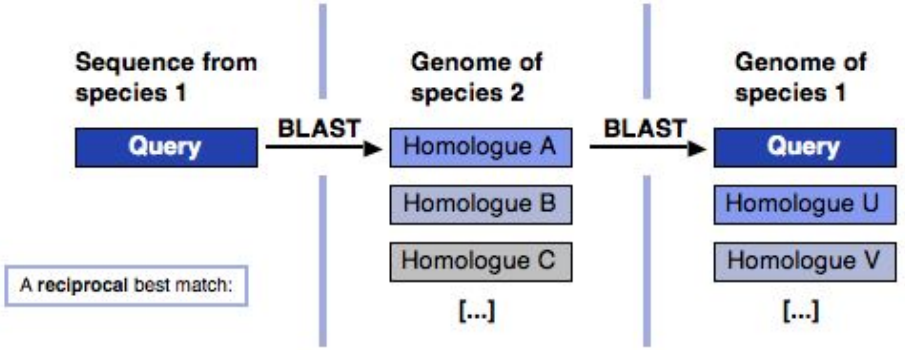
Reciprocal BLAST validation



RED- top hits

GREEN – within two

BLUE – within three



More similar topology and COG label structure

Less similar topology and COG label structure

Baylor College of Medicine

Conclusions

- Consistency radius is a measure of relatedness of protein pairs
  - 30-50% of our "most likely" protein pairs are truly novel orthologs!
  - Protein interaction network and COG self-consistency together predict sequence similarity
- Speculation: this is because important functional networks of proteins are preserved in evolution
  - Maybe some of our protein pairs that don't have similar sequences are functionally similar?
  - Maybe they play similar roles in different pathways?

Next steps

- Further validation
  - Finish processing all seven species we have data about
  - Retrospective analyses... StringDB 9.1 is a year out of date
  - Can we predict what was discovered over the past year?
- Sheaves **seem** natural to transfer information about model organisms, but are they actually effective?
  - Extend processing to other metadata about the proteins in our network
  - Drug interactions, diseases, and pathway networks (BioCyc repository, for instance)

For more information

Michael Robinson

michaelr@american.edu

Preprints available from my website:

http://www.drmichaelrobinson.net/