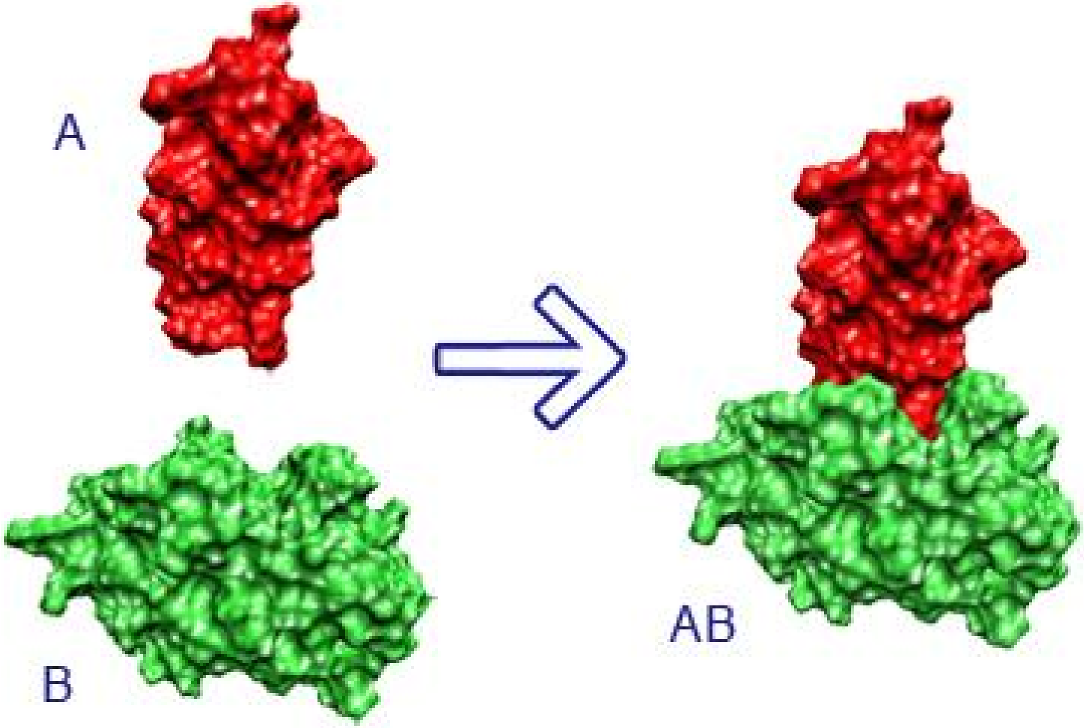
Structure-based prediction of protein-protein interactions on a genome-wide scale.

Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng

Structure-based prediction of protein-protein interactions on a genome-wide scale.

Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng


A
B
AB

How do we know what we know?

How do we know what we know?

High throughput

techniques

How do we know what we know?

High throughput techniques

Computational methods based on non-structural evidence

Structure-based prediction of protein-protein interactions on a genome-wide scale.

Qiangfeng Cliff Zhang

Structure-based pre n of protein-protein interactions on a genome-wide scale.

Homology Models

Qiangfeng Cliff Zhang

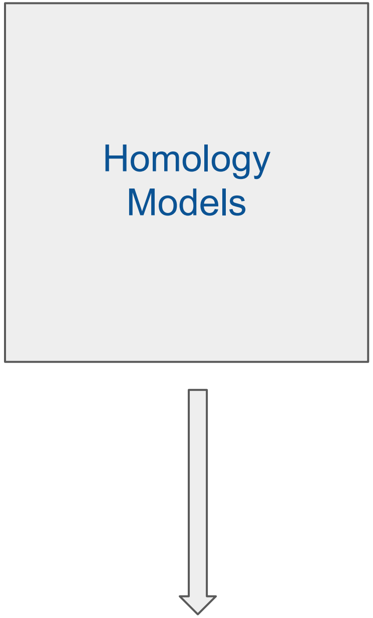
Structure-based pre n of protein-protein interactions on a genome-wide scale.

Homology Models

Structural Neighbors

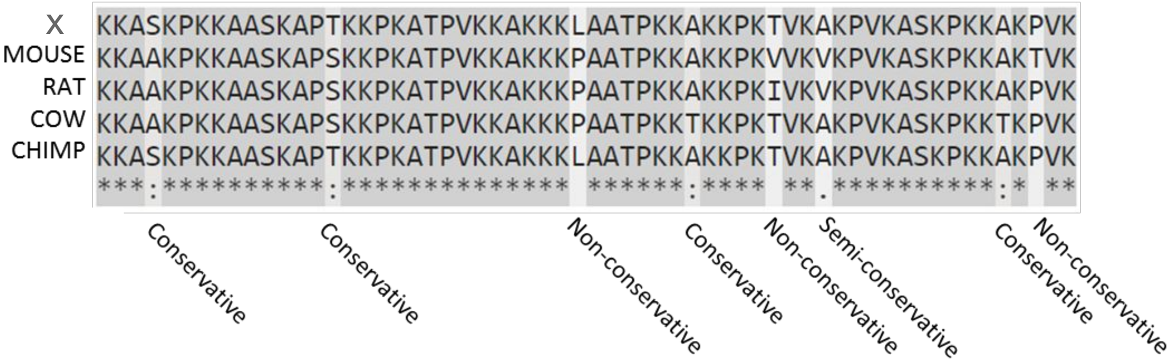
Qiangfeng Cliff Zhang

Homology Models

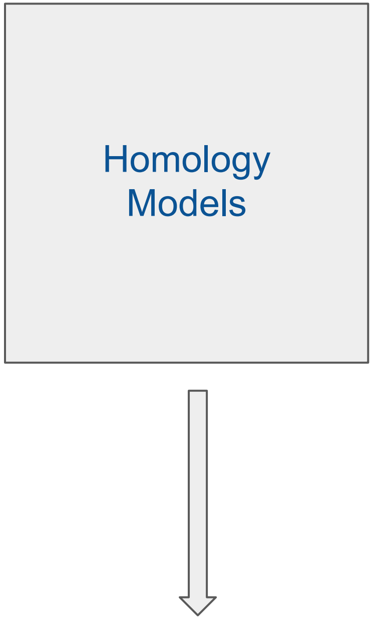


Homology Models

Sequence Alignment



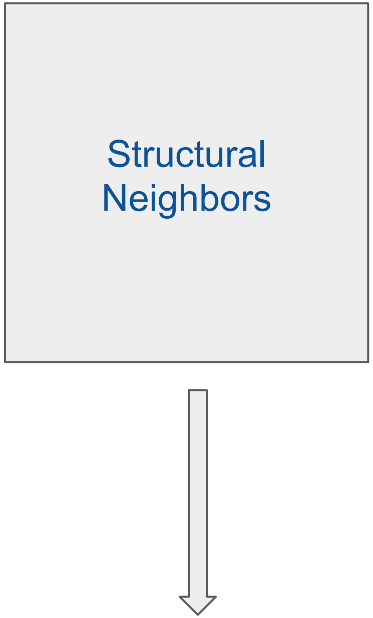
```
X     KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK
RAT   KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKAKPVK
COW   KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
CHIMP KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
      ***:**********:************** ******:**** **.**********:* **
```



Homology Models

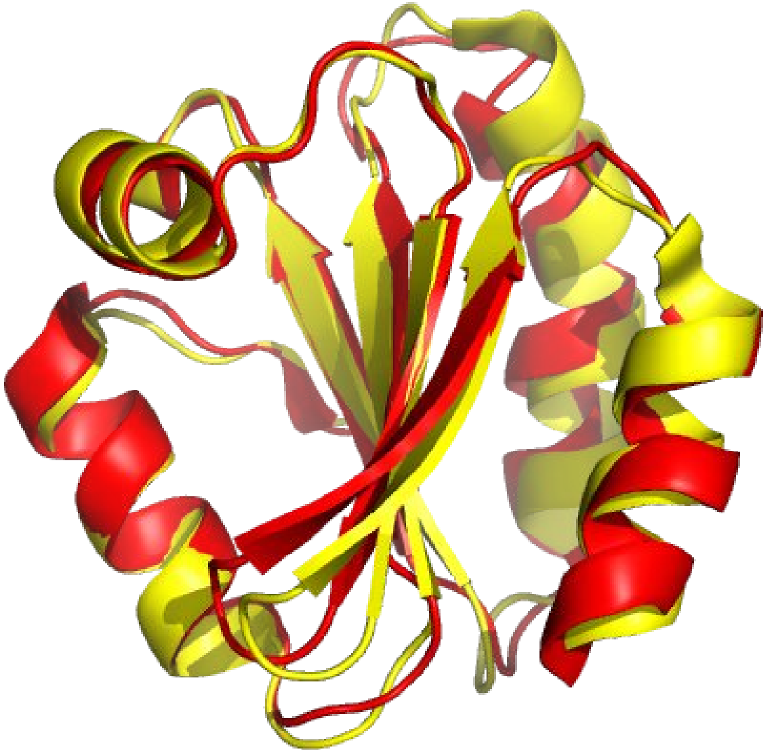
Sequence Alignment

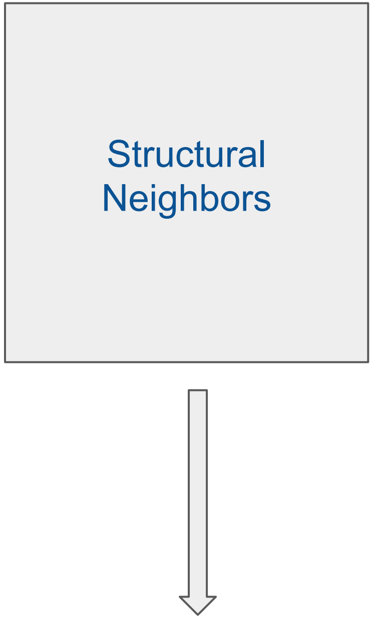
Structural
Neighbors



Structural Neighbors

Structural Alignment





Structural Neighbors

Structural Alignment

Structural information
+
Non-structural information

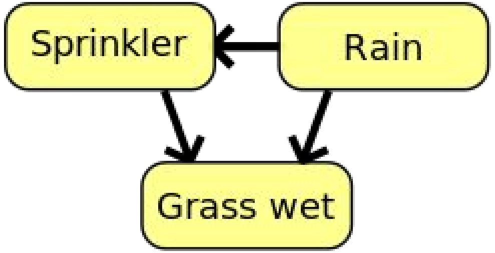
Structural information

+

Non-structural information

=Pre-PPI

NAIVE
BAYESIAN
NETWORK



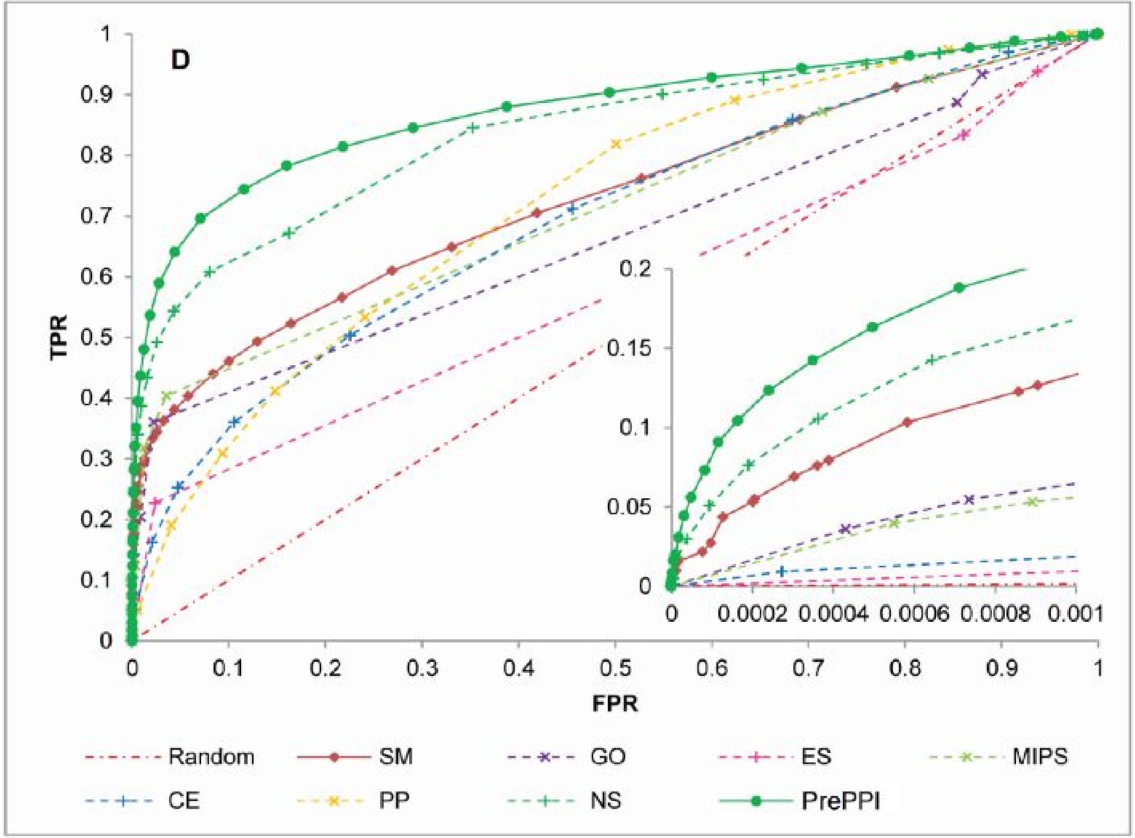
How good is it?

How good is it?

Pre-PPI

VS.

Structural/
Non-structural info.
alone

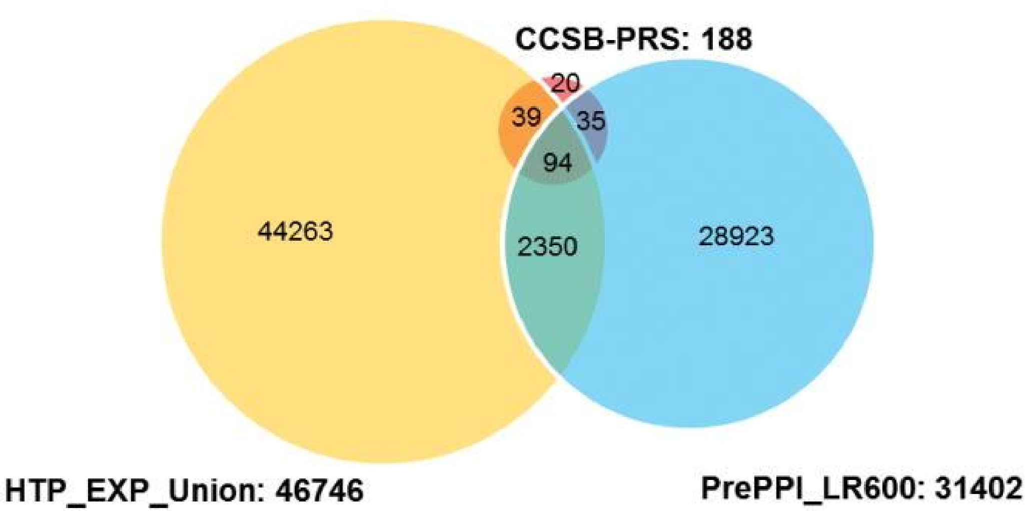


How good is it?

Pre-PPI

VS.

High Throughput Techniques



Panorama of ancient metazoan macromolecular complexes

A paper with two primary authors
and too many others to list

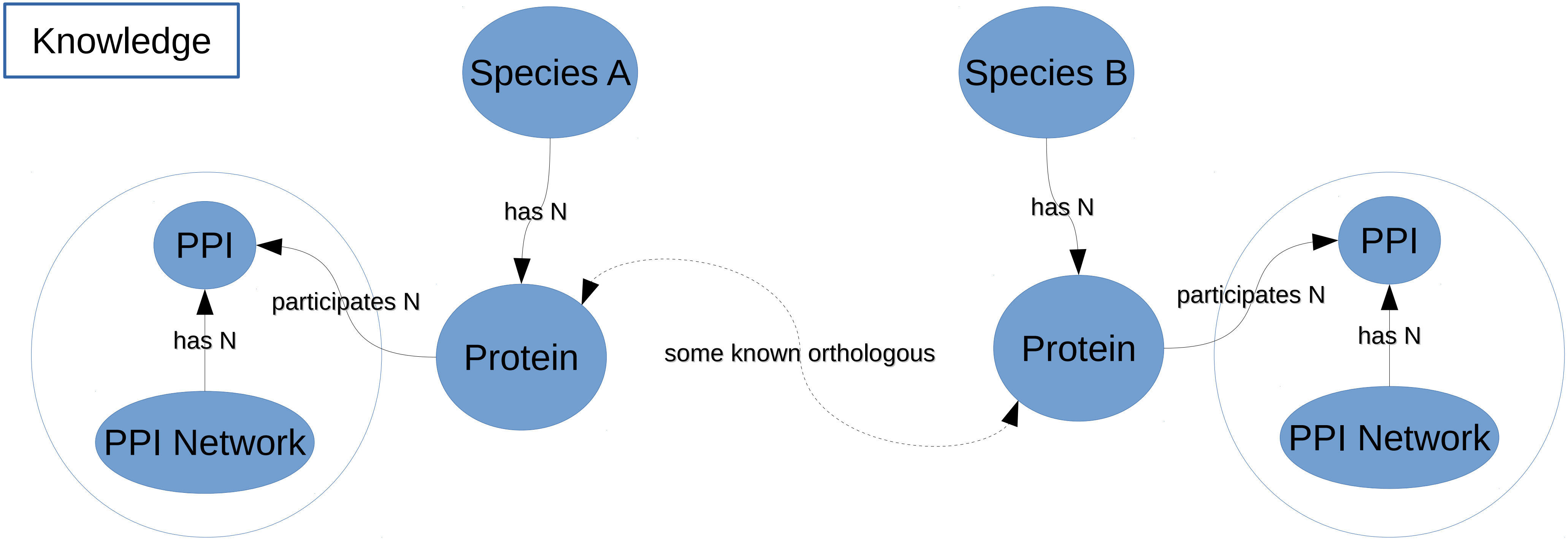
Who should care?

folks who want to

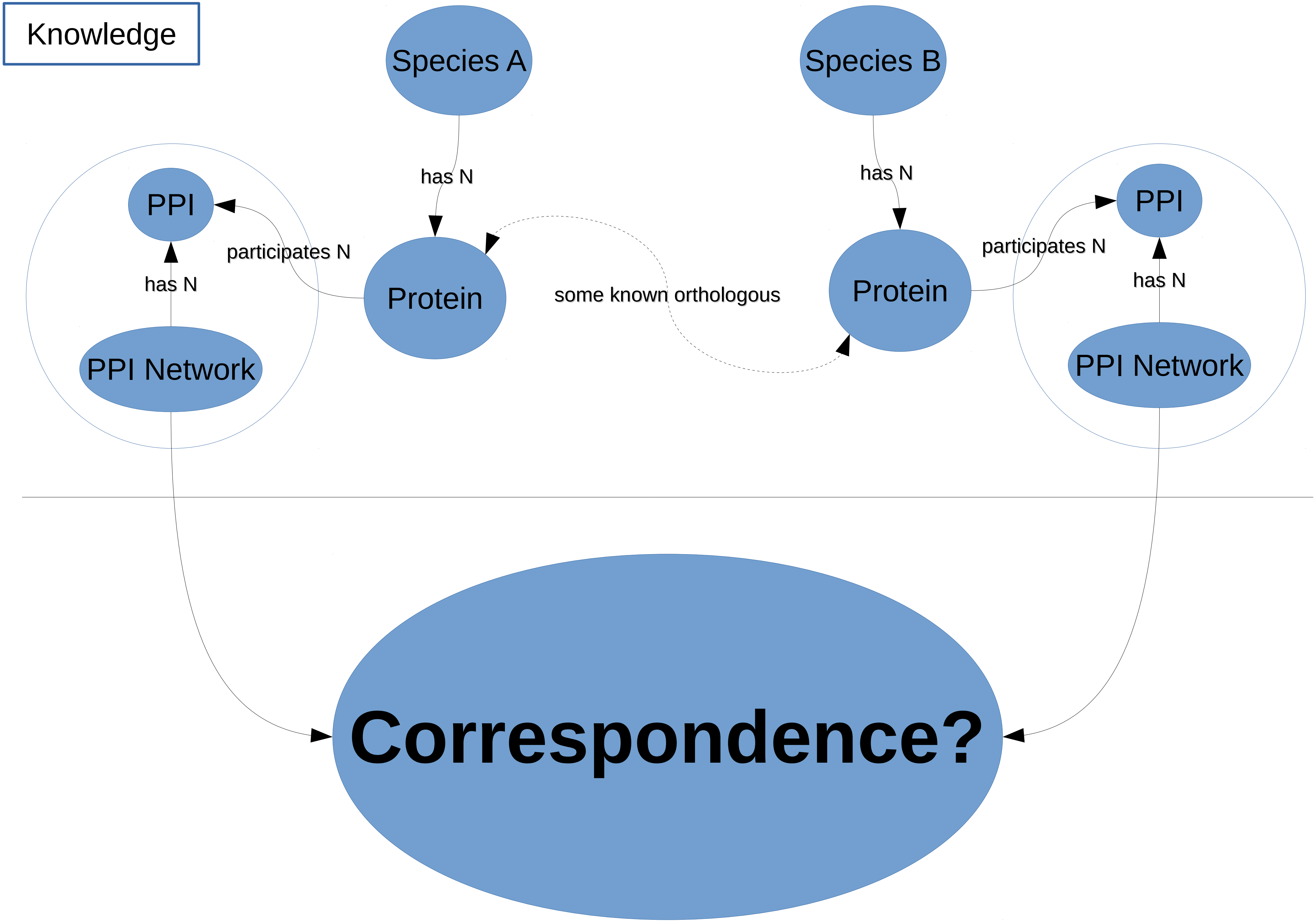
understand

cellular processes

Knowledge


Species A
Species B
has N
has N
PPI
PPI
participates N
participates N
has N
has N
Protein
Protein
some known orthologous
PPI Network
PPI Network

Knowledge



Species A

Species B

has N

has N

PPI

PPI

participates N

participates N

has N

has N

Protein

Protein

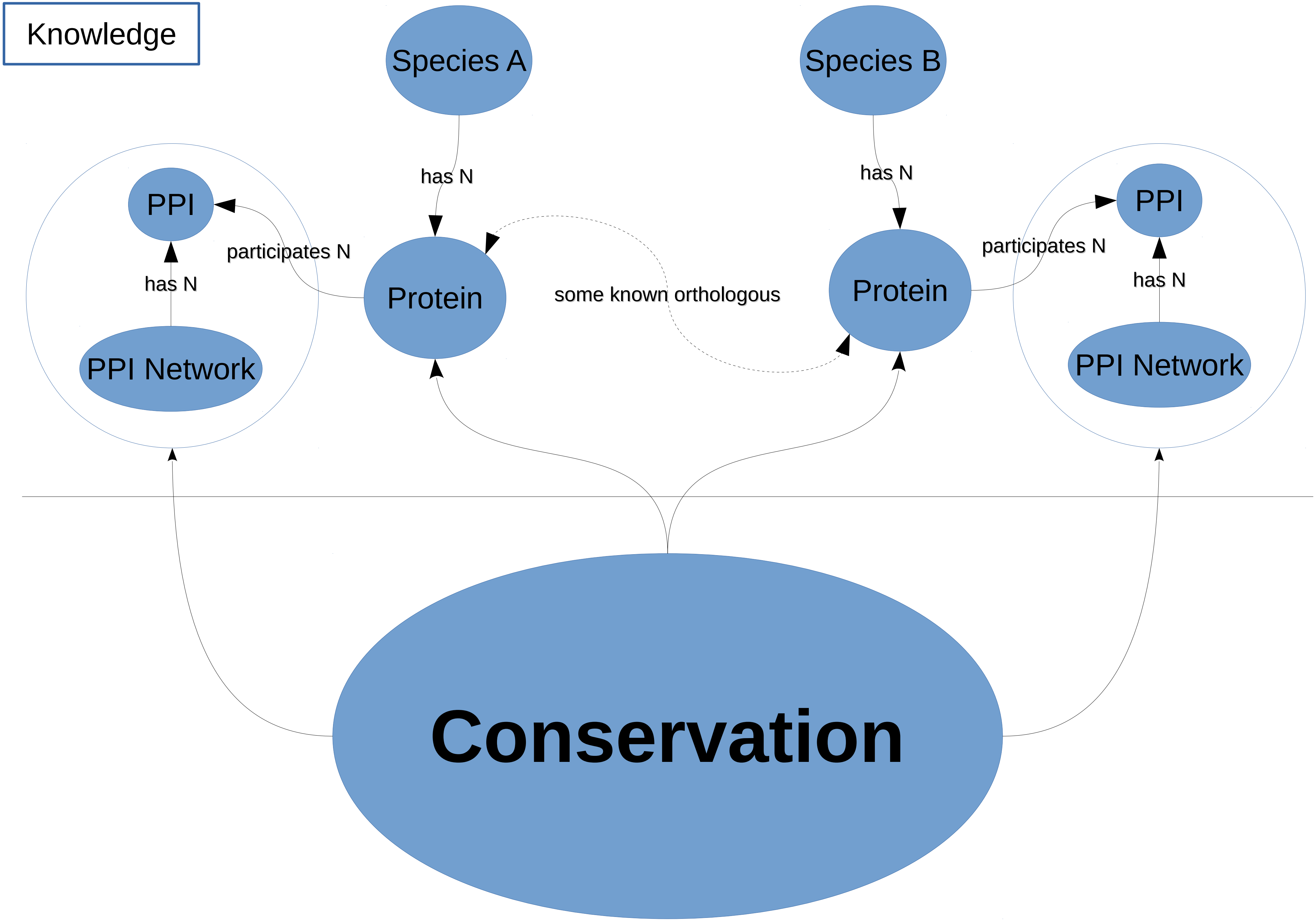
some known orthologous

PPI Network

PPI Network

Correspondence?

Knowledge



Species A

Species B

has N

has N

PPI

participates N

Protein

some known orthologous

Protein

participates N

PPI

has N

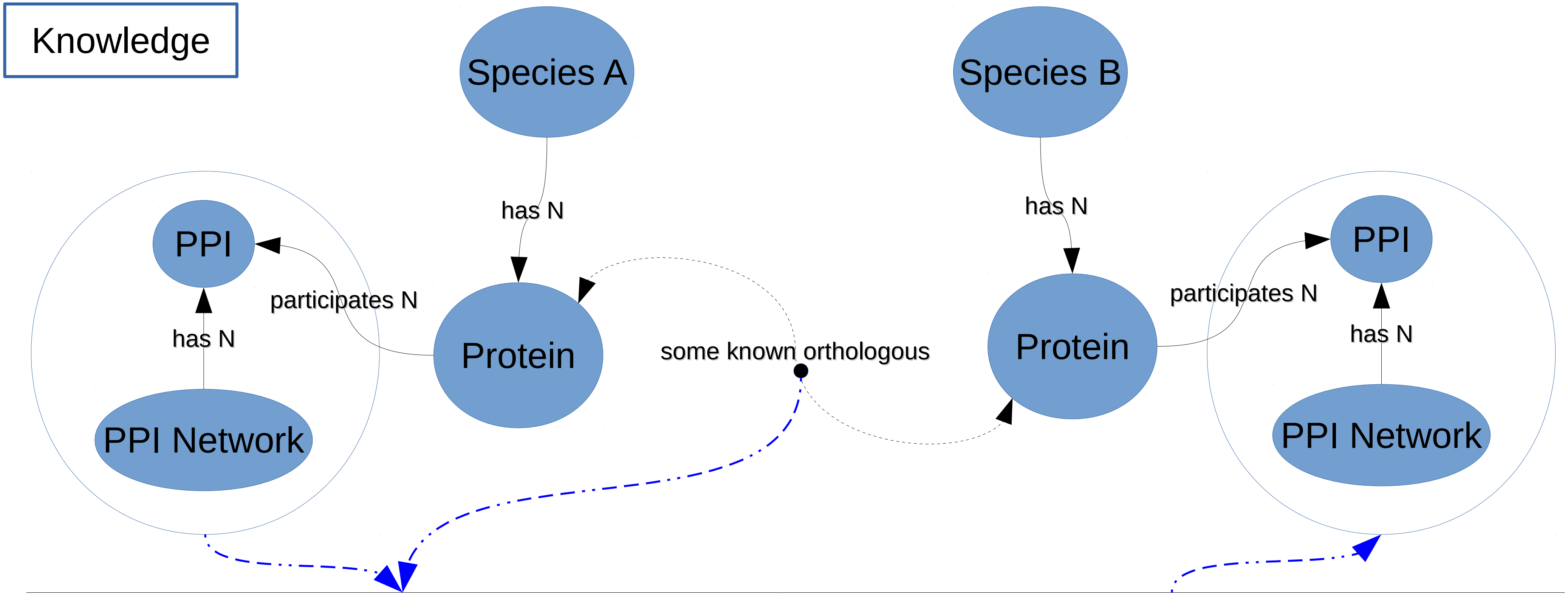
has N

PPI Network

PPI Network

Conservation

Knowledge



Methods

Experimental

- Co-fractionation
- Mass Spectrometry
- *stuff out of presentation scope*

Computational

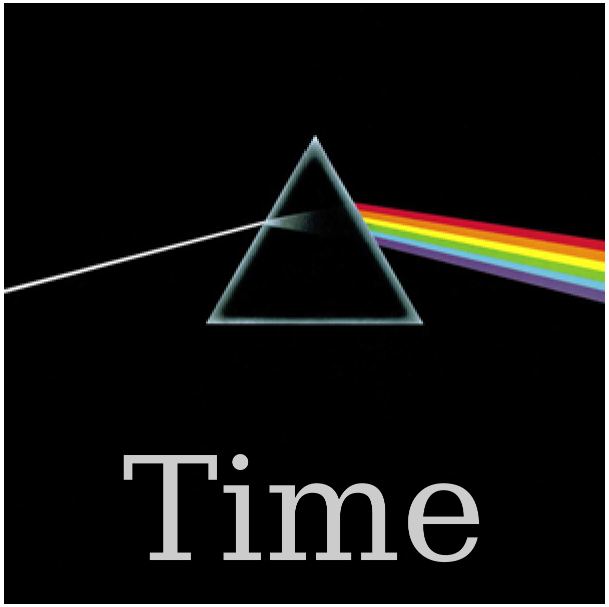
- Mapping
- Machine Learning
- Graph Clustering
- Enrichment Analysis
- *stuff out of presentation scope*

10,000ft Objective

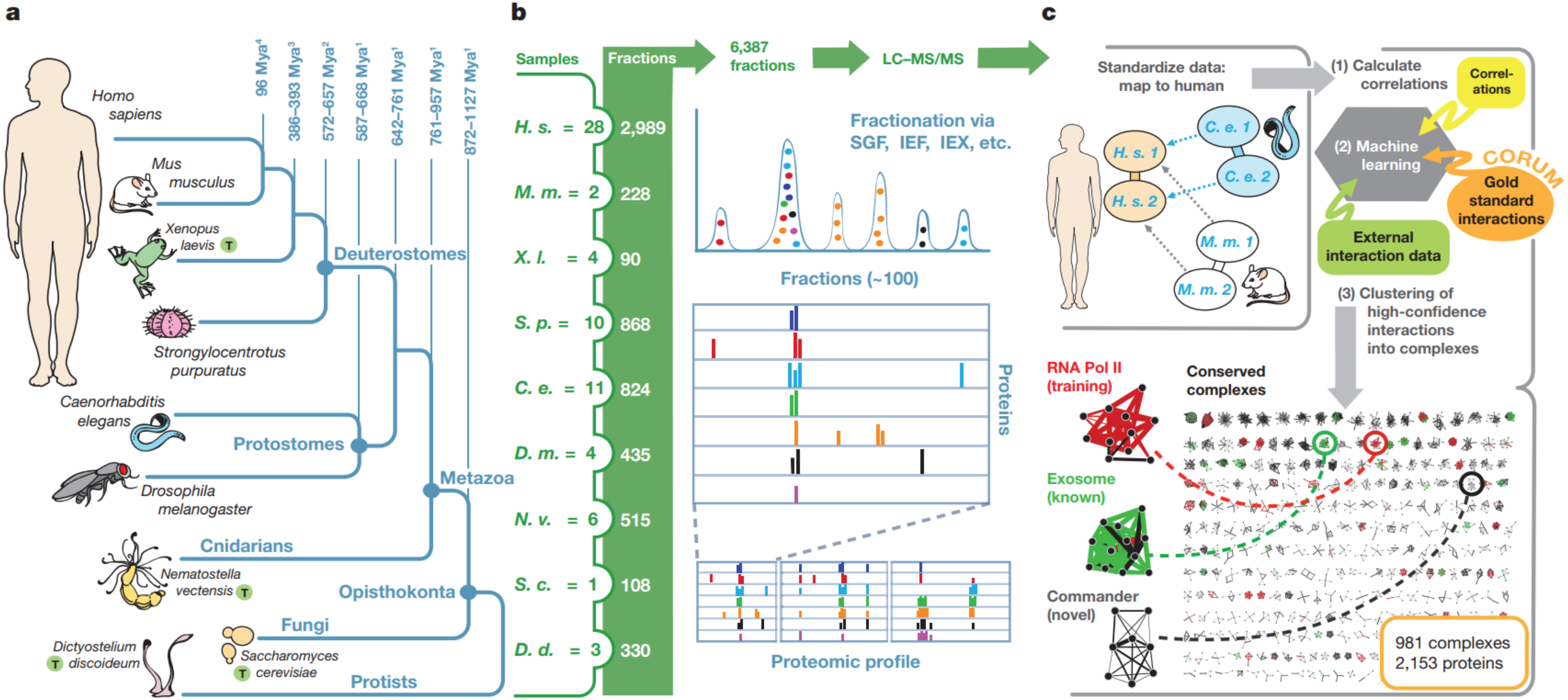
- Describe extent of PPI network conservation
- Identify conserved PPI networks
- Characterize conserved PPI networks

10,000ft Objective

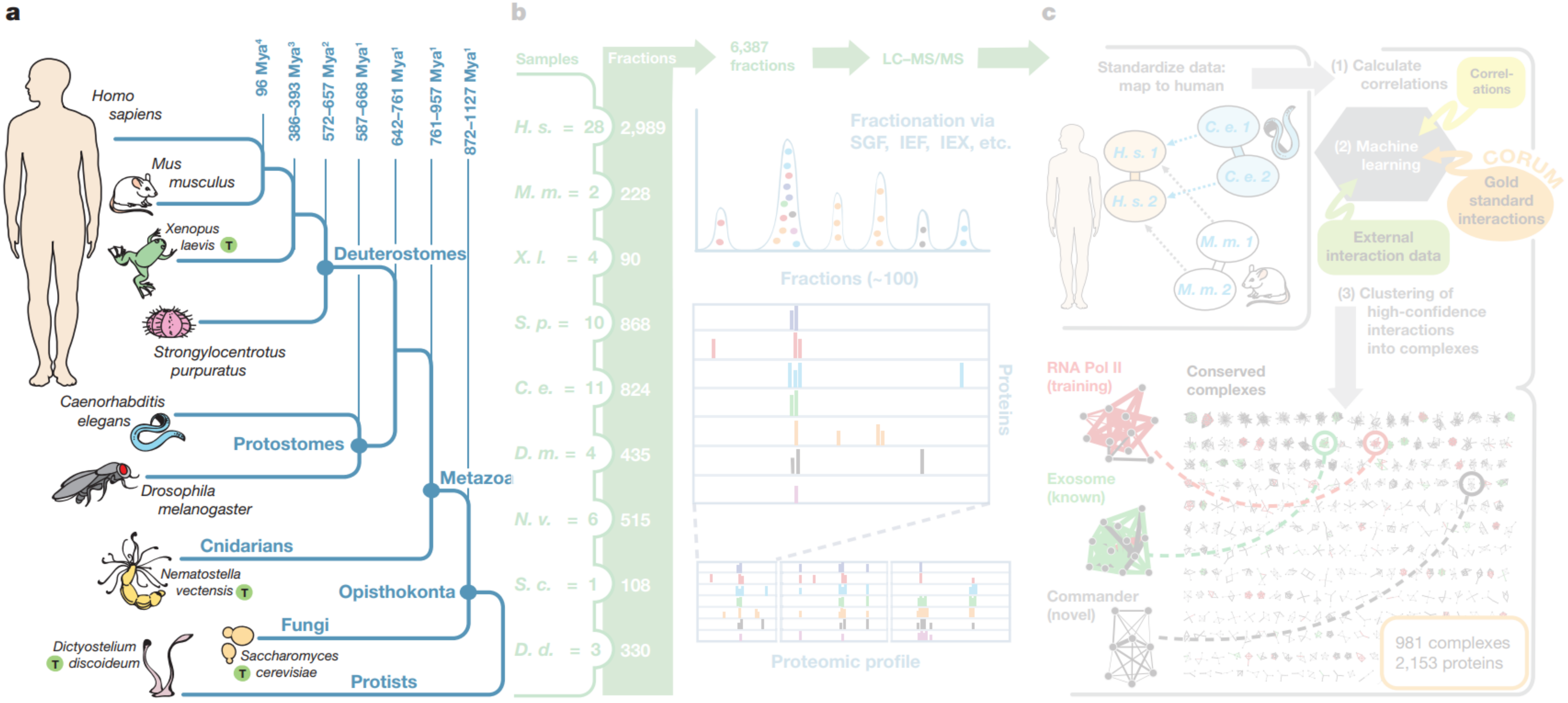
- Describe extent of PPI network conservation
- Identify conserved PPI networks
- Characterize conserved PPI networks



Methods: Overview



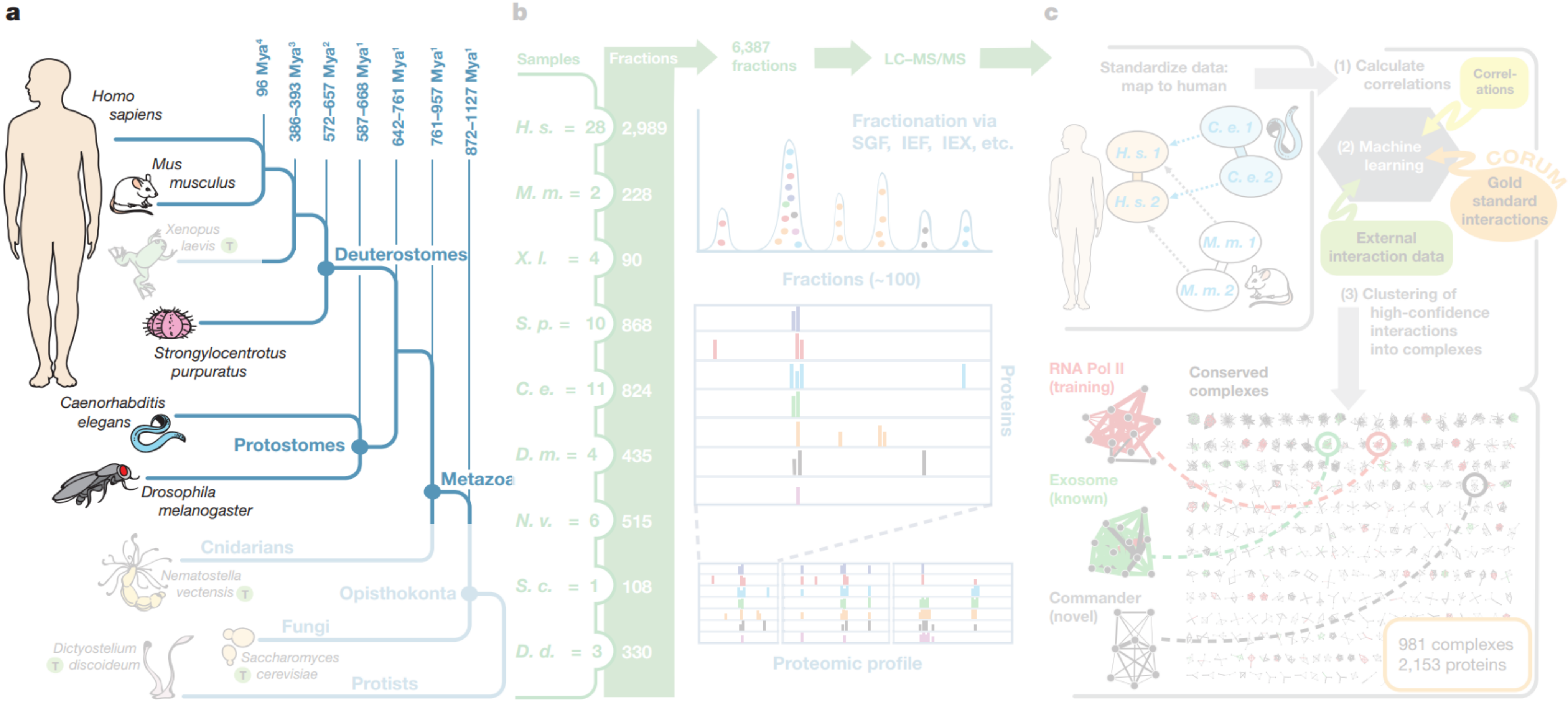
Population



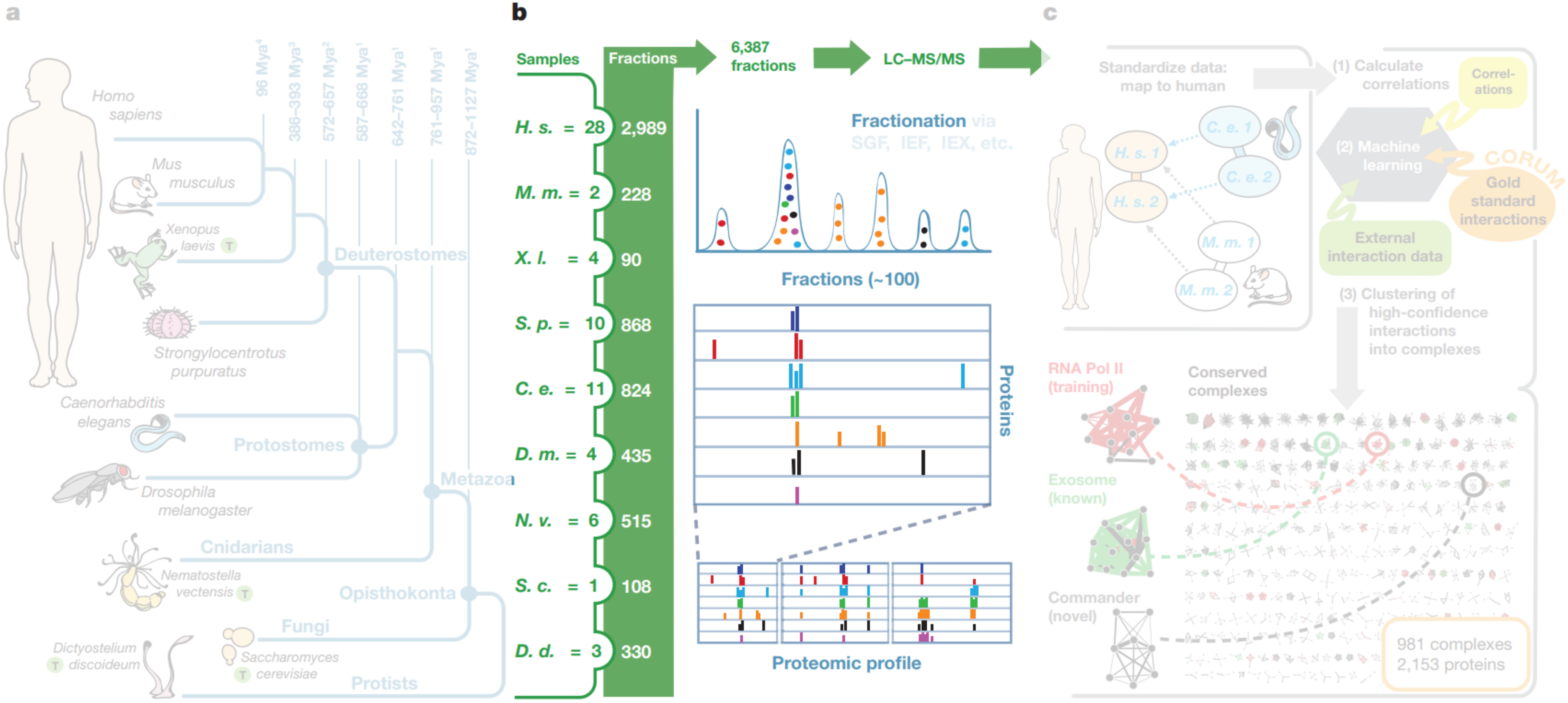
a
96 Mya4
386–393 Mya3
572–657 Mya2
587–668 Mya1
642–761 Mya1
761–957 Mya1
872–1127 Mya1
Homo sapiens
Mus musculus
Xenopus laevis T
Strongylocentrotus purpuratus
Deuterostomes
Caenorhabditis elegans
Protostomes
Drosophila melanogaster
Metazoa
Cnidarians
Nematostella vectensis T
Opisthokonta
Fungi
Saccharomyces T cerevisiae
Dictyostelium T discoideum
Protists
b
Samples
Fractions
6,387 fractions
LC–MS/MS
H. s. = 28
2,989
M. m. = 2
228
X. l. = 4
90
S. p. = 10
868
C. e. = 11
824
D. m. = 4
435
N. v. = 6
515
S. c. = 1
108
D. d. = 3
330
Fractionation via SGF, IEF, IEX, etc.
Fractions (~100)
Proteins
Proteomic profile
c
Standardize data: map to human
H. s. 1
H. s. 2
C. e. 1
C. e. 2
M. m. 1
M. m. 2
(1) Calculate correlations
Correl-ations
(2) Machine learning
CORUM
Gold standard interactions
External interaction data
(3) Clustering of high-confidence interactions into complexes
RNA Pol II (training)
Exosome (known)
Commander (novel)
Conserved complexes
981 complexes
2,153 proteins

Population

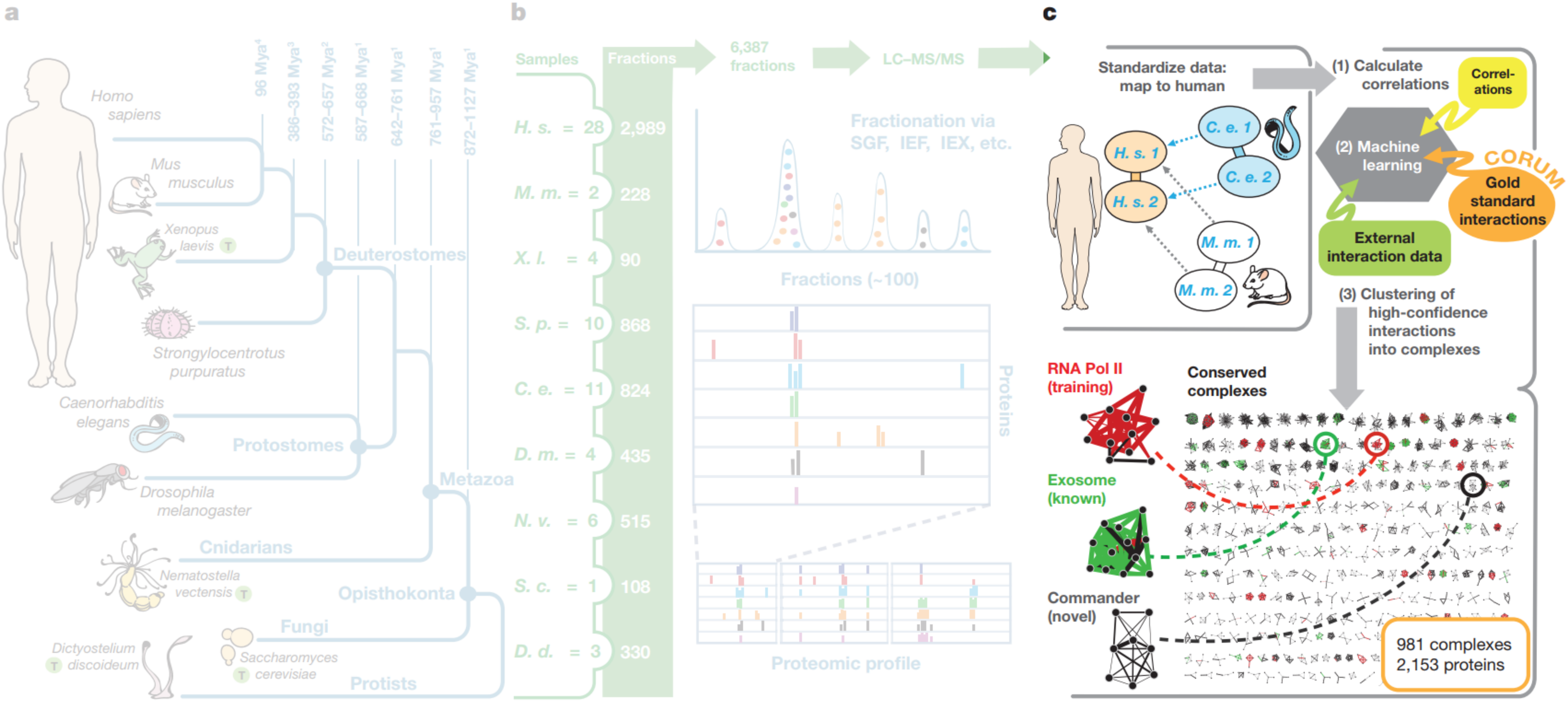
Training



Data Acquisition



Data Analysis



The data analysis is the hard part.

Naturally that’s what we’re gonna focus on.

Basis for Comparison

Have

- Interspecies PPI pairs

Want

- Comparable PPI pairs

Solutions

- PPI pairs over orthogroups
- PPI pairs normalized to a single species
- ???

Basis for Comparison

Have

- Interspecies PPI pairs

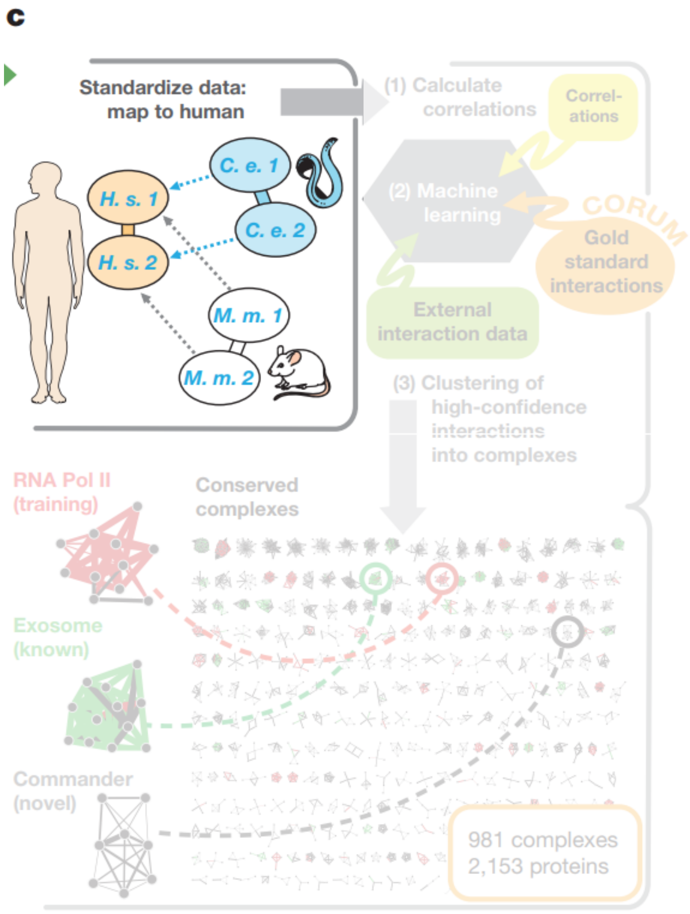
Want

- Comparable PPI pairs

Solutions

- PPI pairs over orthogroups
- PPI pairs normalized to a single species *(human)*
- ???

arbitrary



Pause for a thought

Have

- Comparable PPI pairs w/ co-fractionation data

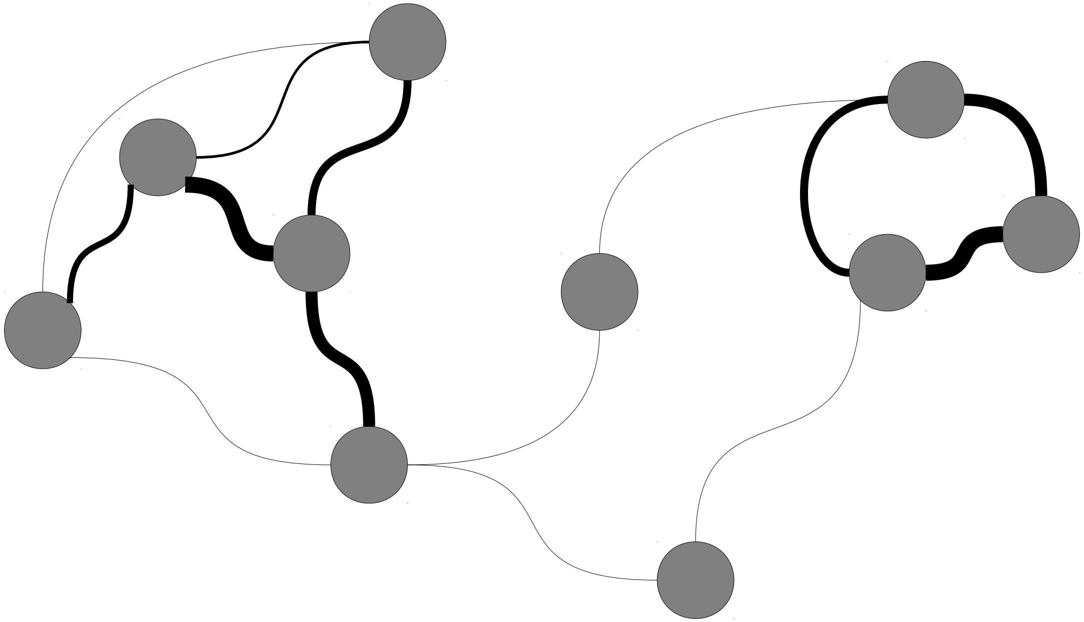
Want

- Real PPI networks

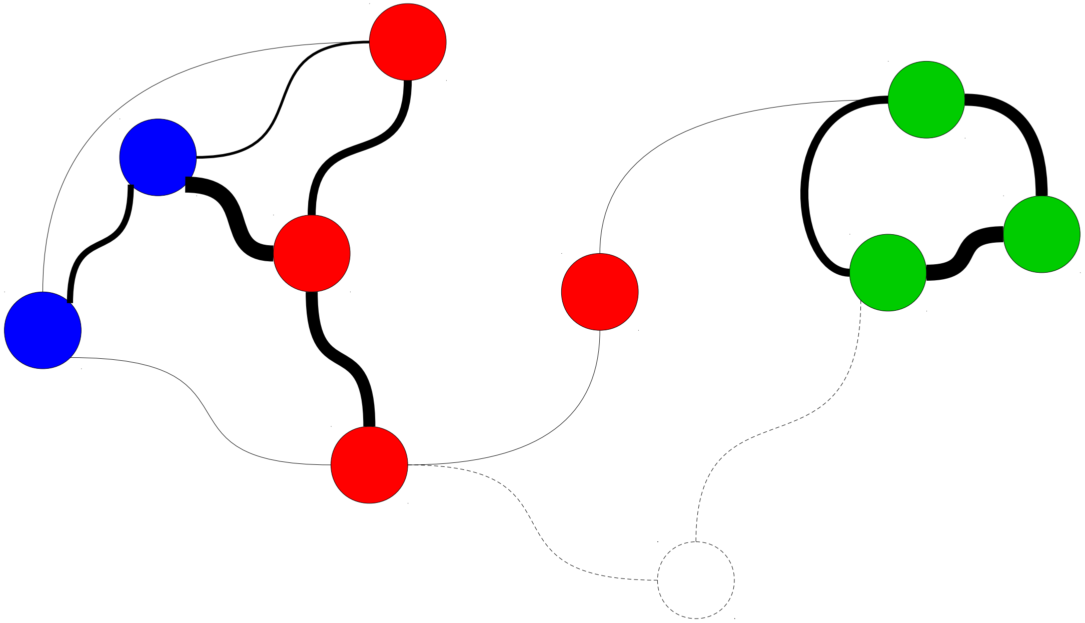
A Solution

- Score PPI pairs for being co-complex
  - Informed by gold-standard adjacencies
- Transform PPI pairs into a global adjacency network
- Split out subnetworks from global PPI network
  - Tuned for meaningful PPI networks

Graph Clustering Primer

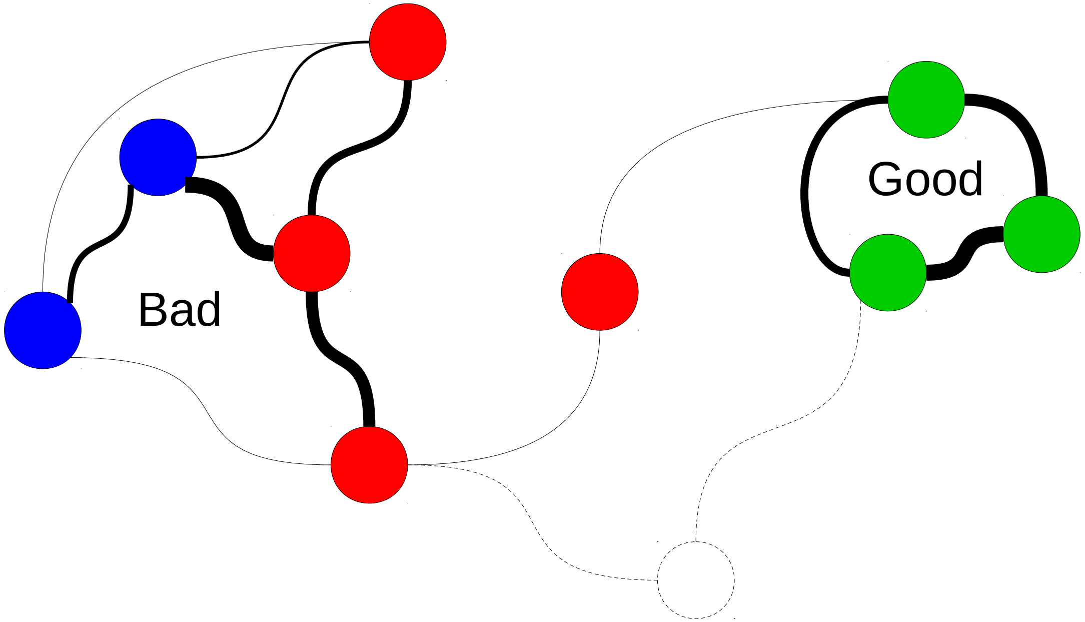


Graph Clustering Primer

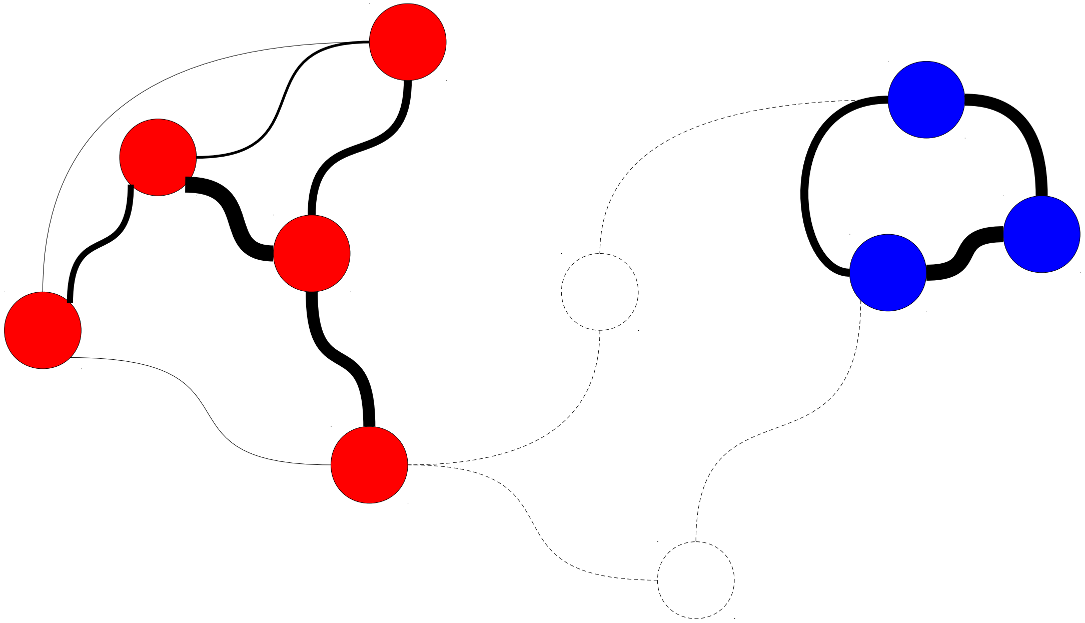


example of a ‘clustering’

Graph Clustering Primer



Graph Clustering Primer

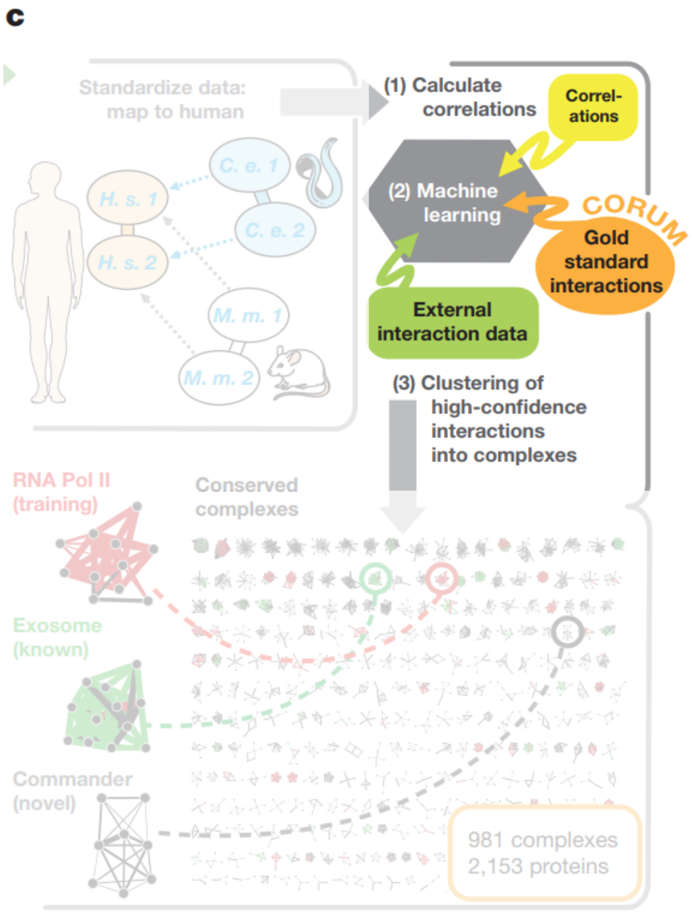


probably want something like this

We now return to your scheduled programming

Gluing together a data pipeline

- Generate features
  - Co-fractionation correlation measures
  - Interaction data from databases, literature
- ML co-complex +/- against CORUM
- Network clustering w/ ClusterOne, MCL



Characterization of Clusters

Have

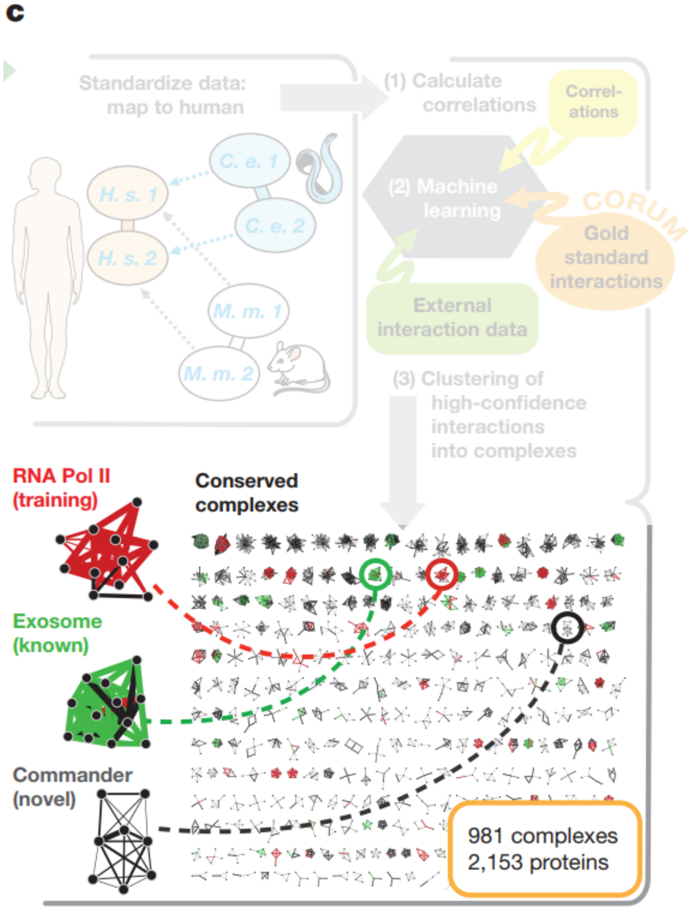
- PPI networks embedded in global network

Want

- Biologically relevant and ‘interesting’ features

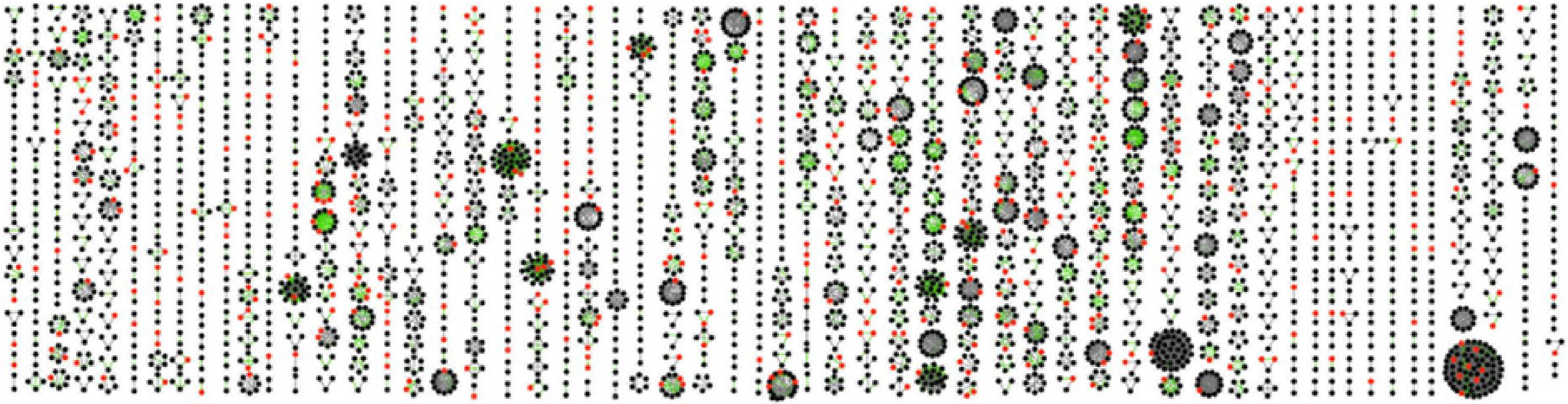
A Solution

- Enrichment analysis for higher-order attributes



Smattering of Results

found many neglected proteins/PPIs of interest

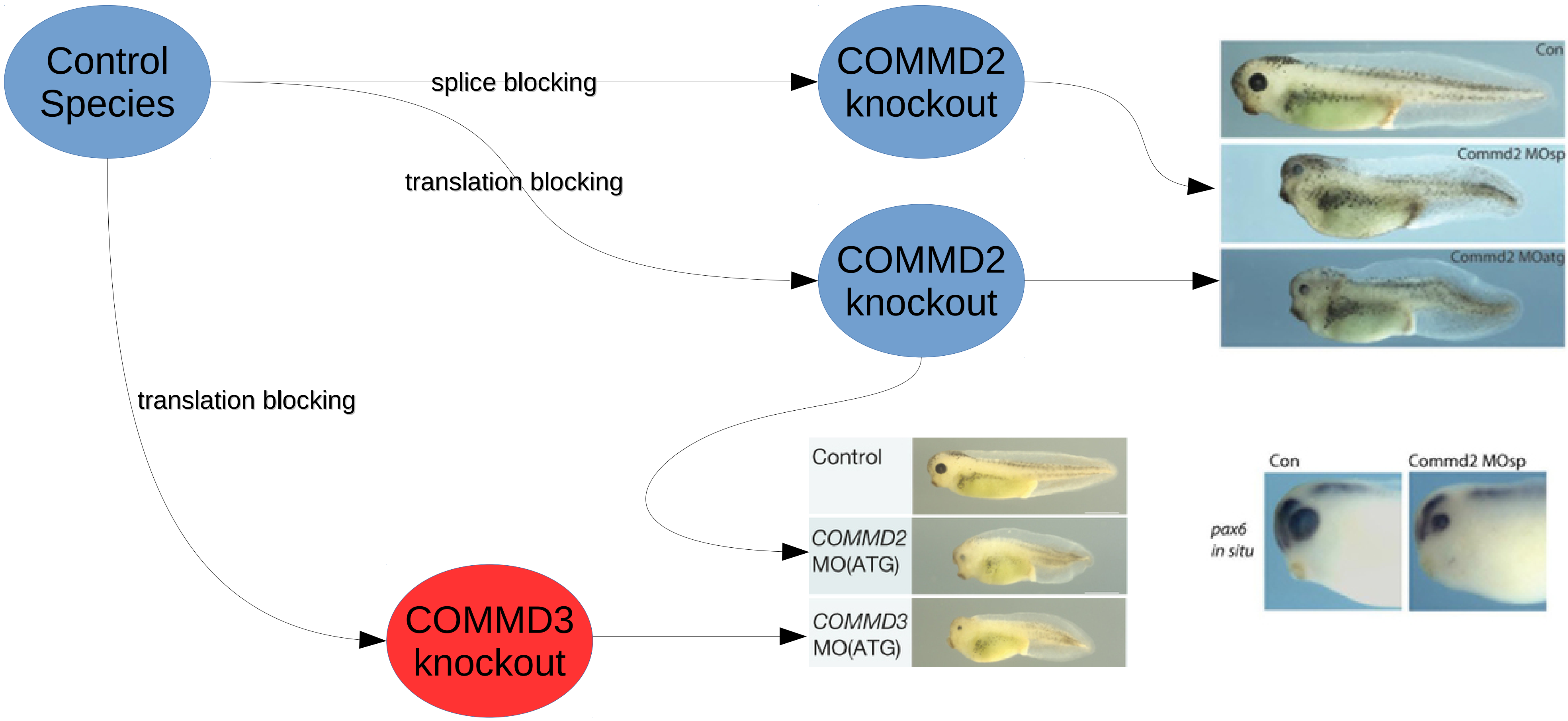


green links are novel co-complex interactions

red dots are unannotated proteins

Smattering of Results

directly demonstrated utility of approach



Control Species

splice blocking

COMMD2 knockout

translation blocking

COMMD2 knockout

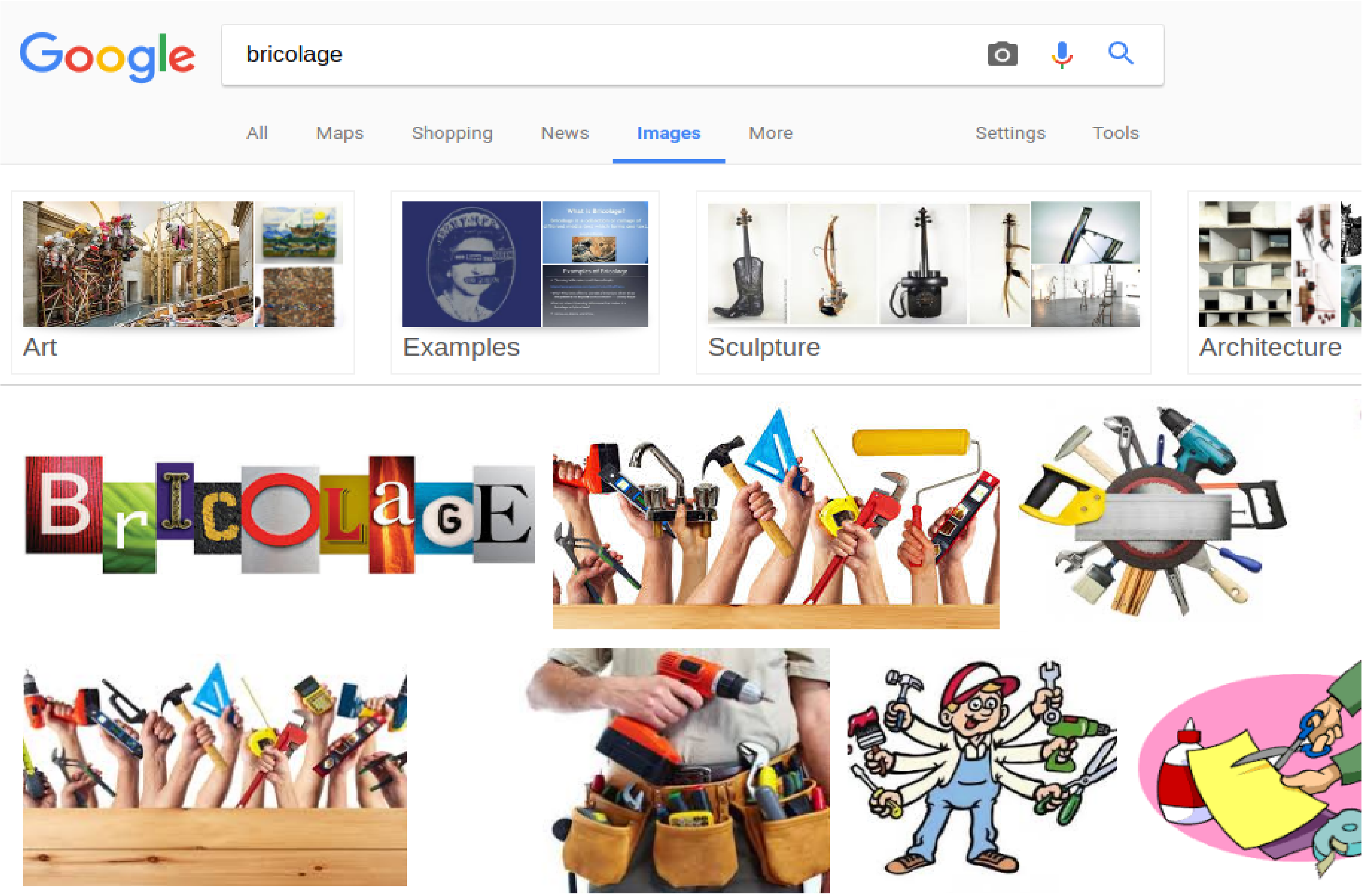
translation blocking

COMMD3 knockout

Things I wish the authors mentioned

- Were multiple co-fractionation correlation measures really necessary? Which features were relevant?
- Did they account for the false-discovery rate of clusters?
- Was there a specific reason for choosing humans as the ortholog basis? How much do results change when using a different species?

Final Comments



bricolage

The candy is a lie