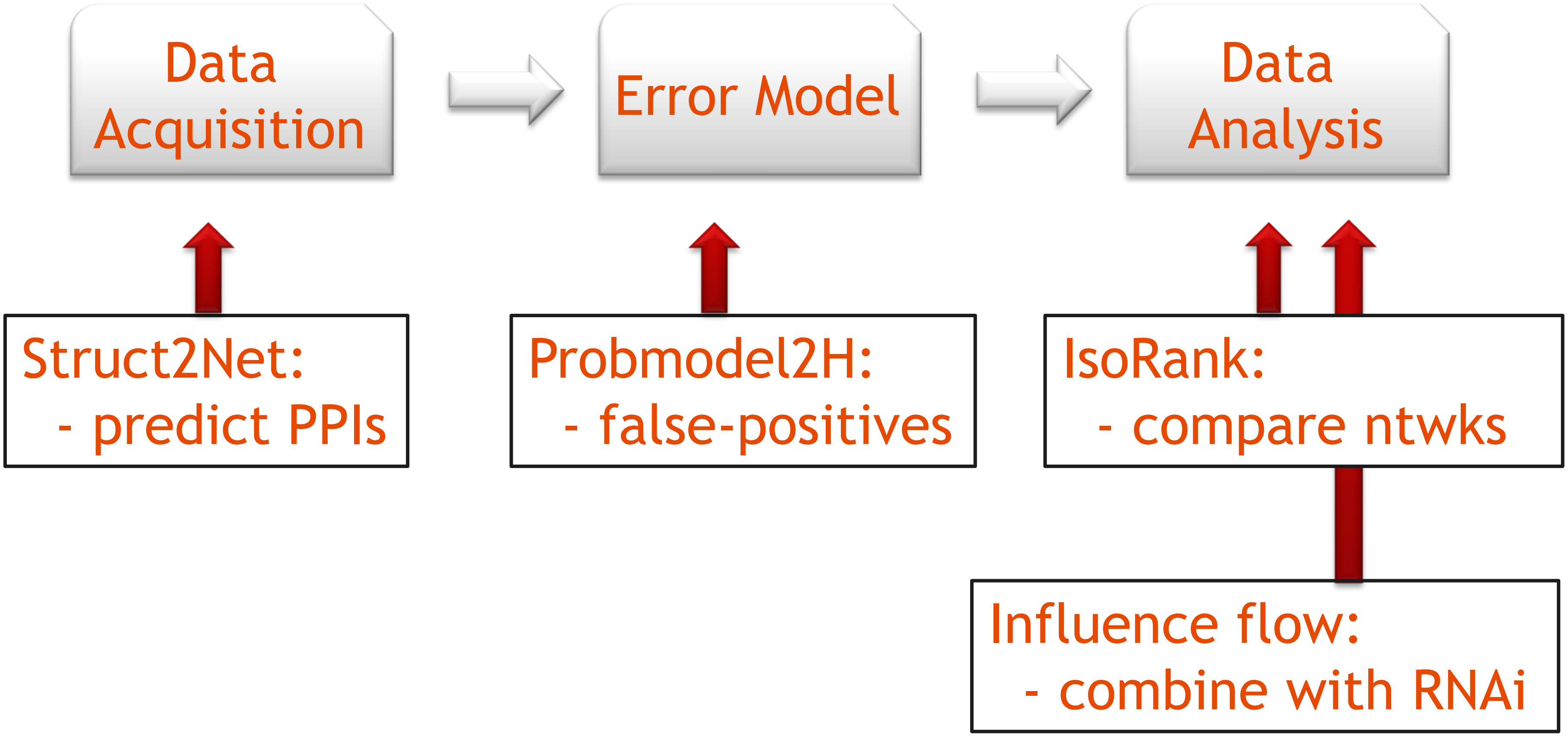
Algorithms for the Analysis of Protein Interaction Networks

Rohit Singh
MIT

Thesis Defense
July 27, 2011

Outline

- Introduction to Protein Interactions
- Algorithms for PPI Networks:



Data Acquisition

Error Model

Data Analysis

Struct2Net:
 - predict PPIs

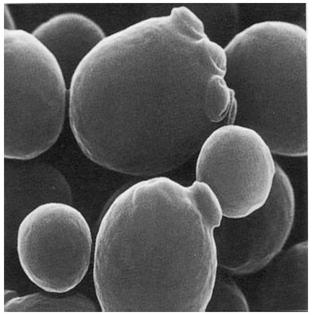
Probmodel2H:
 - false-positives

IsoRank:
 - compare ntwks

Influence flow:
 - combine with RNAi

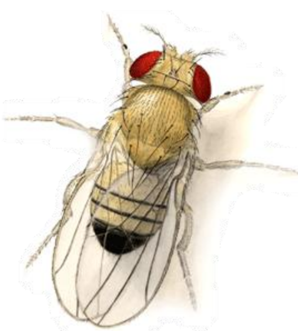
Protein interactions are crucial to the cellular system

- Proteins interact with other proteins to perform their functions
- Many cellular activities are a result of protein interactions

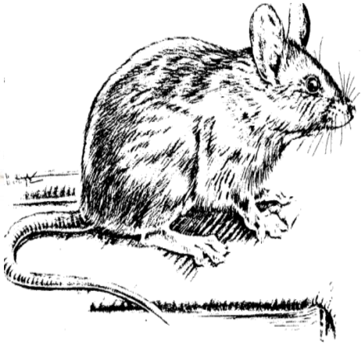


6600

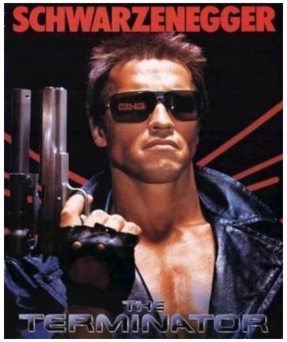
20000



14000



24500



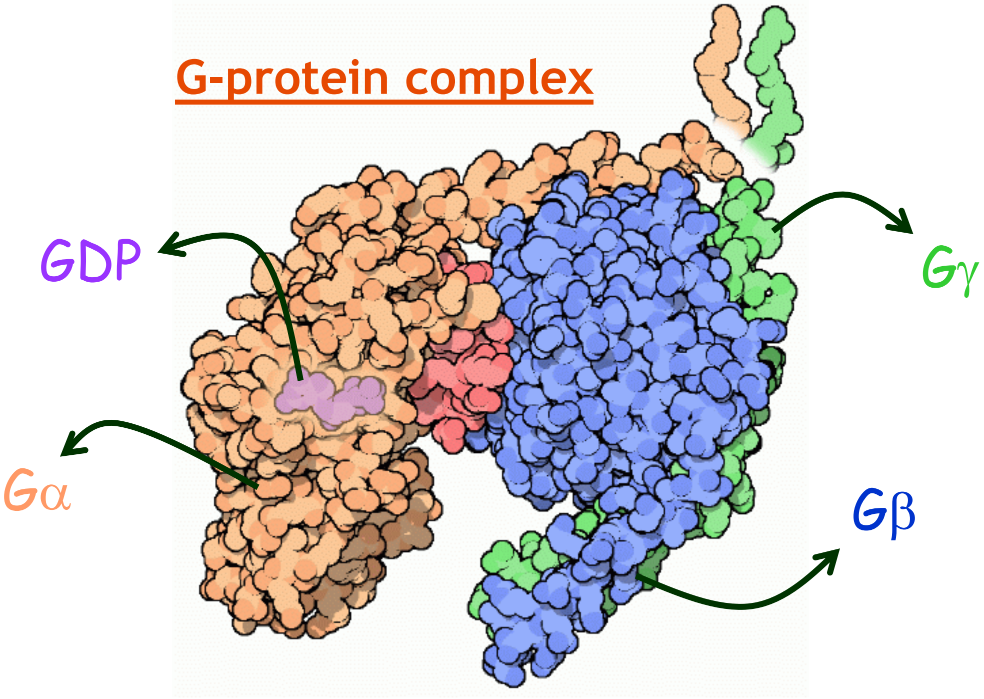
23000

Number of Genes

Numbers from http://www.ensembl.org

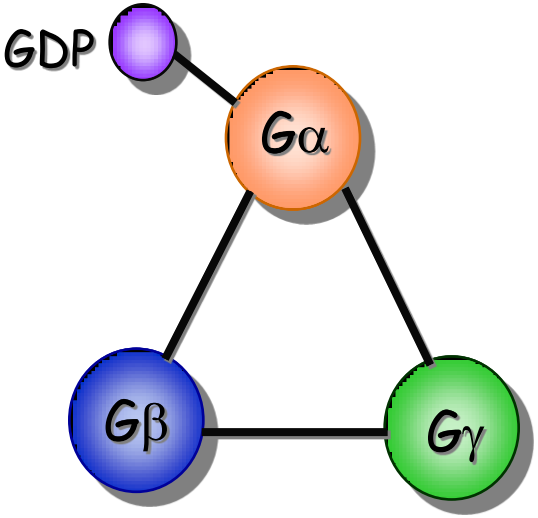
In recent years, the approach to PPI analysis has changed

- Old perspective: low-throughput, structural
- New perspective: high-throughput, graph-based



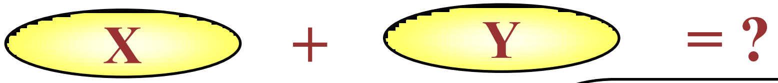
Old perspective

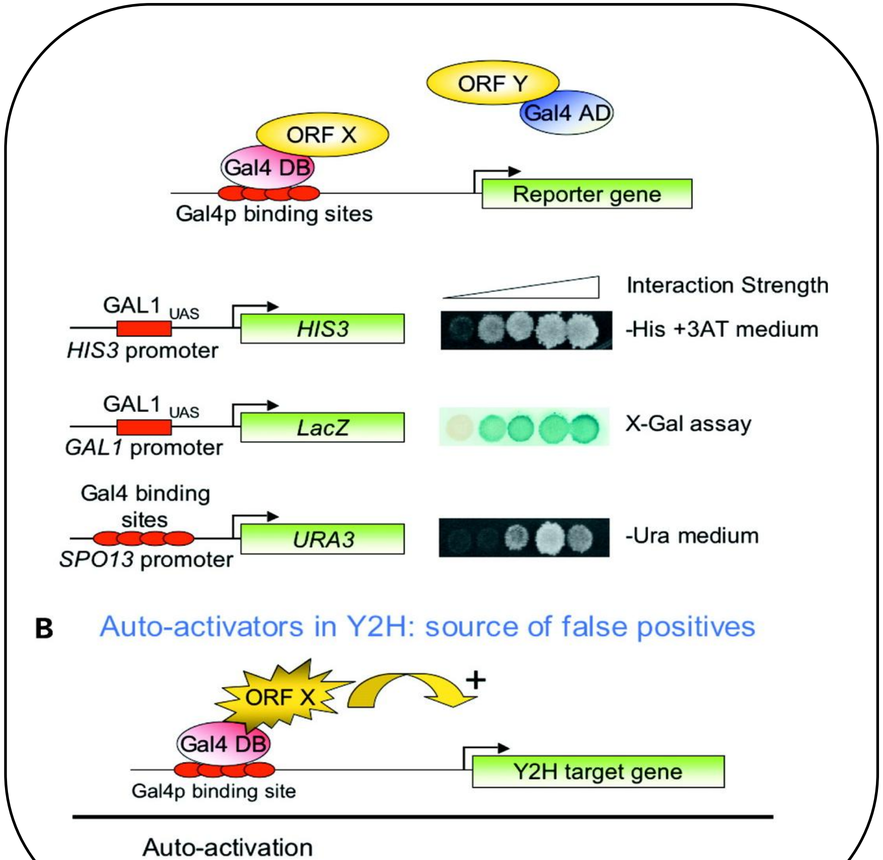
Image from www.rcsb.org



New perspective

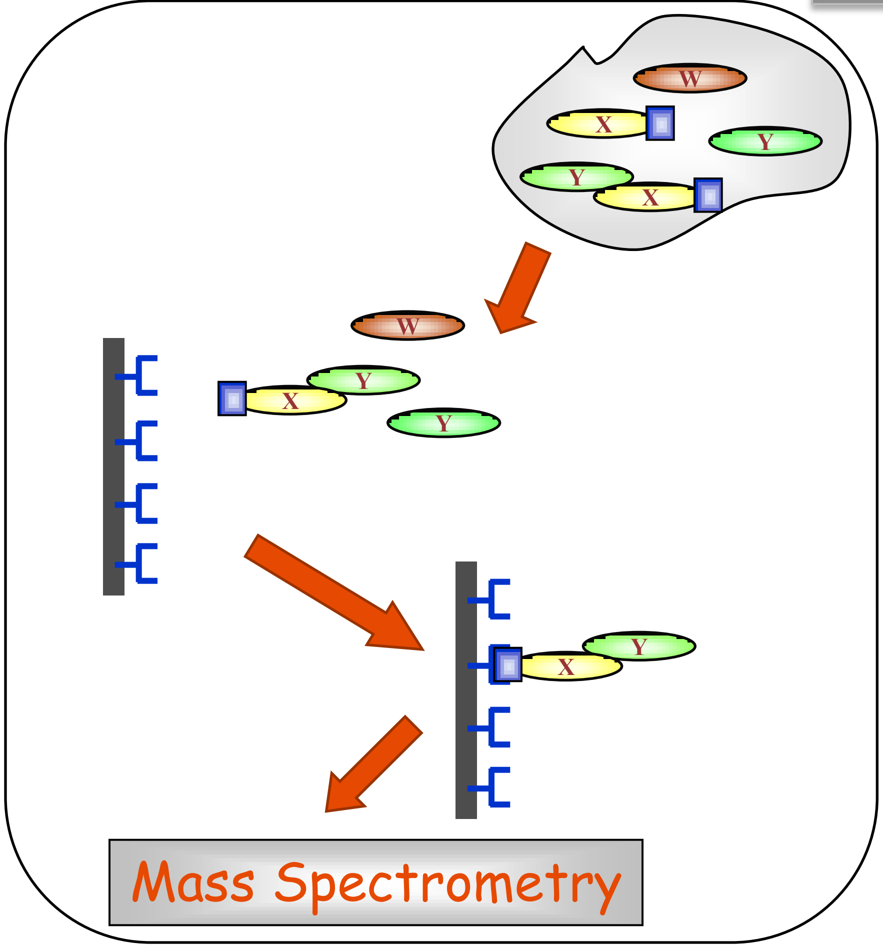
High-throughput experiments are providing a lot of PPI data…





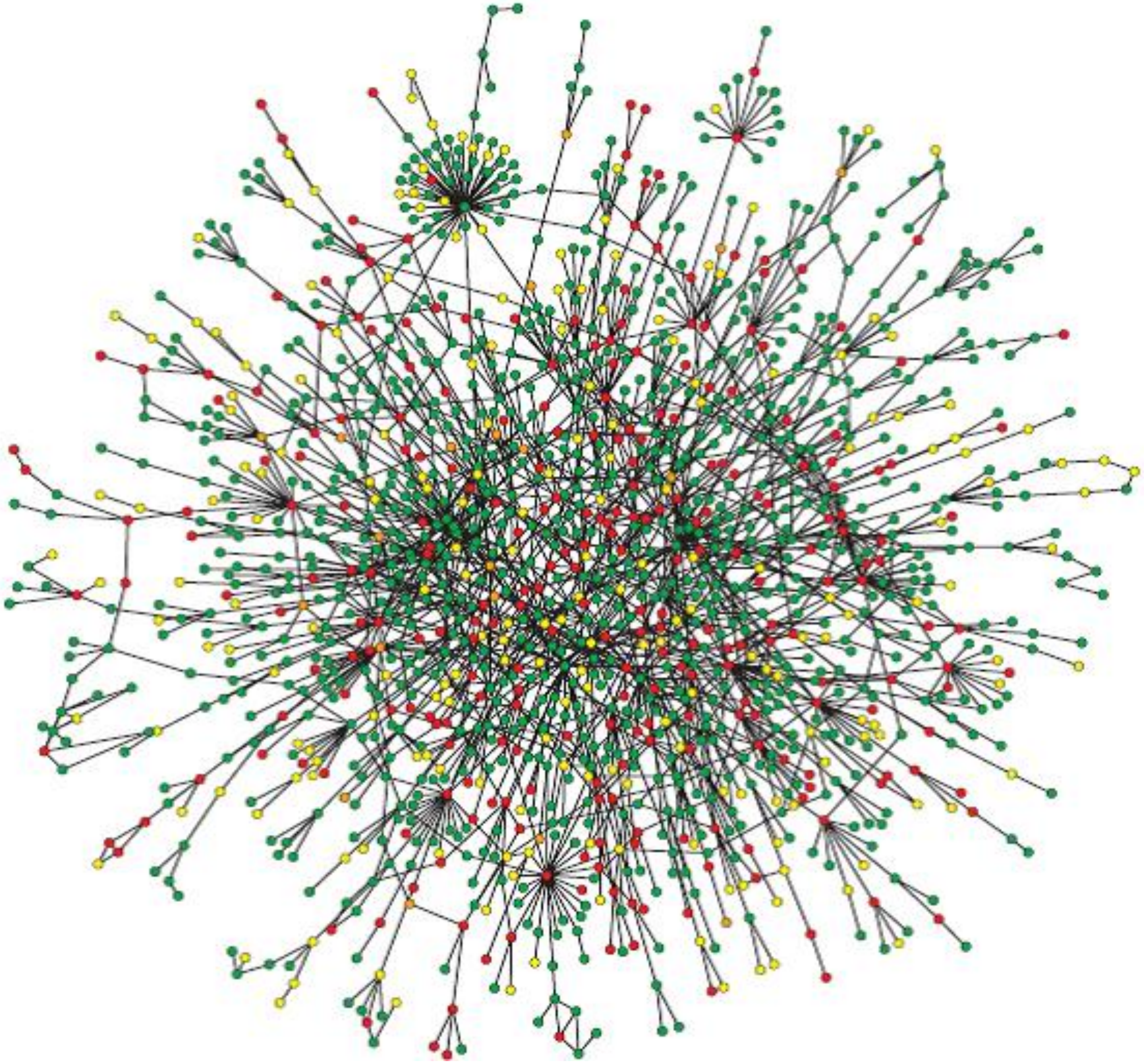
Cusick et al. Hum Med Gen, 05

Yeast Two-Hybrid



Co-immunoprecipitation

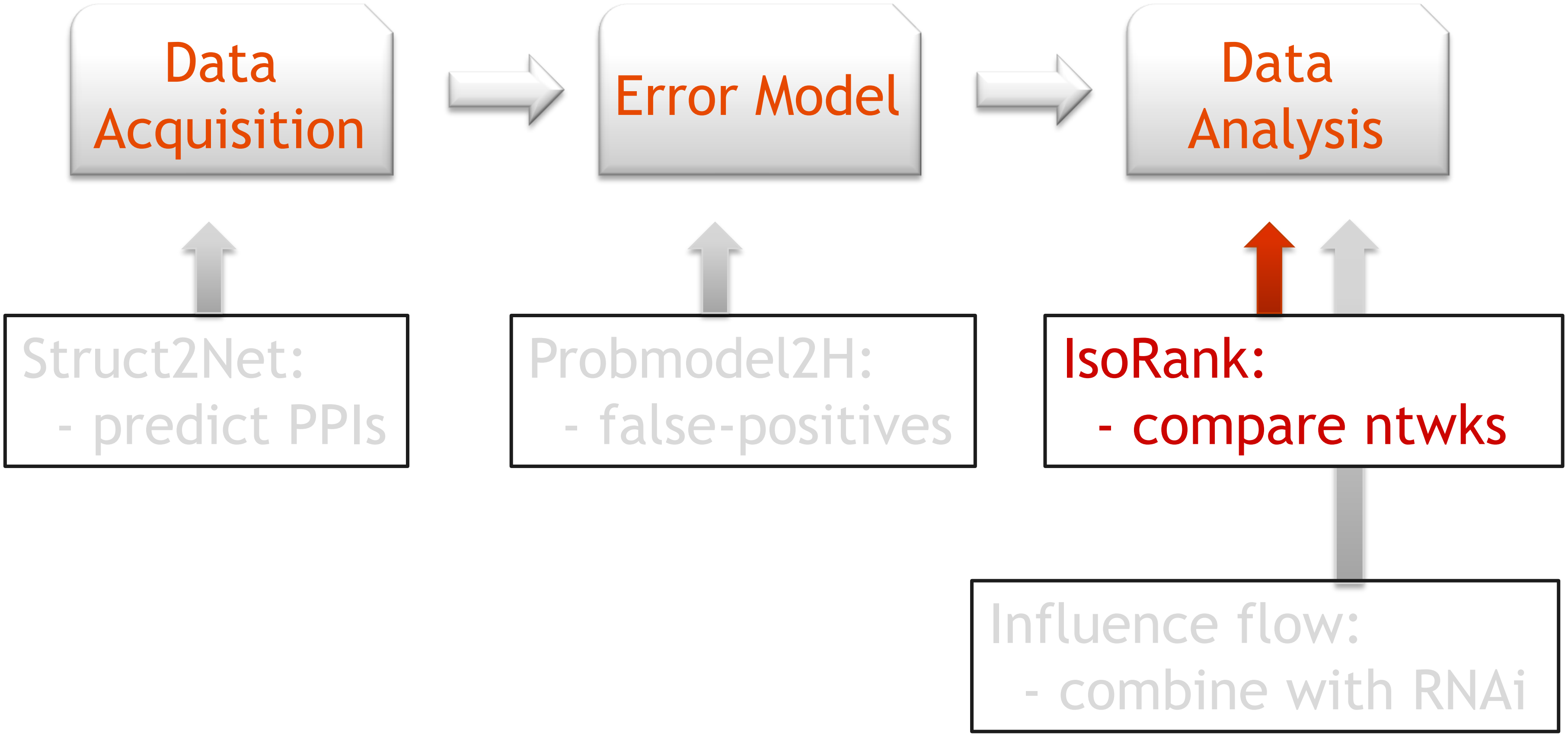
An Example PPI Network: Yeast



http://compbio.pbworks.com/f/1166443065/protein_map.gif

Outline

- Introduction to Protein Interactions
- Algorithms for PPI Networks:



Data Acquisition

Error Model

Data Analysis

Struct2Net:
- predict PPIs

Probmodel2H:
- false-positives

IsoRank:
- compare ntwks

Influence flow:
- combine with RNAi

IsoRank & IsoRankN

Goal: global alignment of PPI networks

Why?

- Comparative genomics on a network level
- Estimate functional orthologs: gene correspondences across species

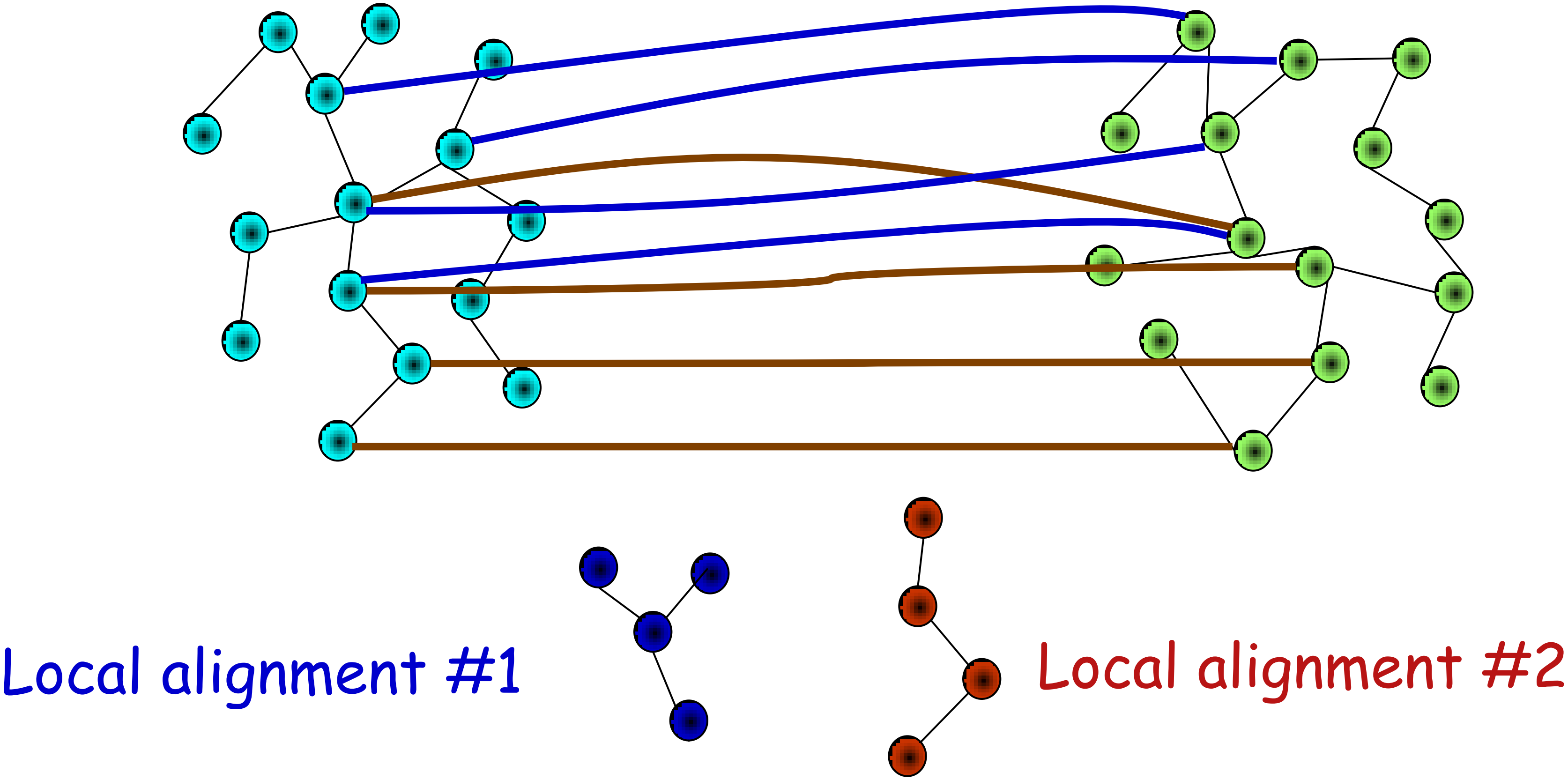
How?

- Intuition: match nodes whose neighborhood topologies match
- Construct an eigenvalue problem

Acknowledgments

- Collaborators:
  - IsoRank: Jinbo Xu & Bonnie Berger
  - IsoRankN: Chung-Shou Liao, Kanghao Lu, Michael Baym & Bonnie Berger
  - IsoBase: Daniel Park, Michael Baym & Bonnie Berger
- Previously presented/published:
  - RECOMB 2007
  - PSB 2008
  - Proceedings of the Nat'l Acad. Of Sciences, 2008
  - ISMB 2009 & BioInformatics 2009
  - Nucleic Acids Research (Database Issue) 2011

Network Alignment: Local vs. Global



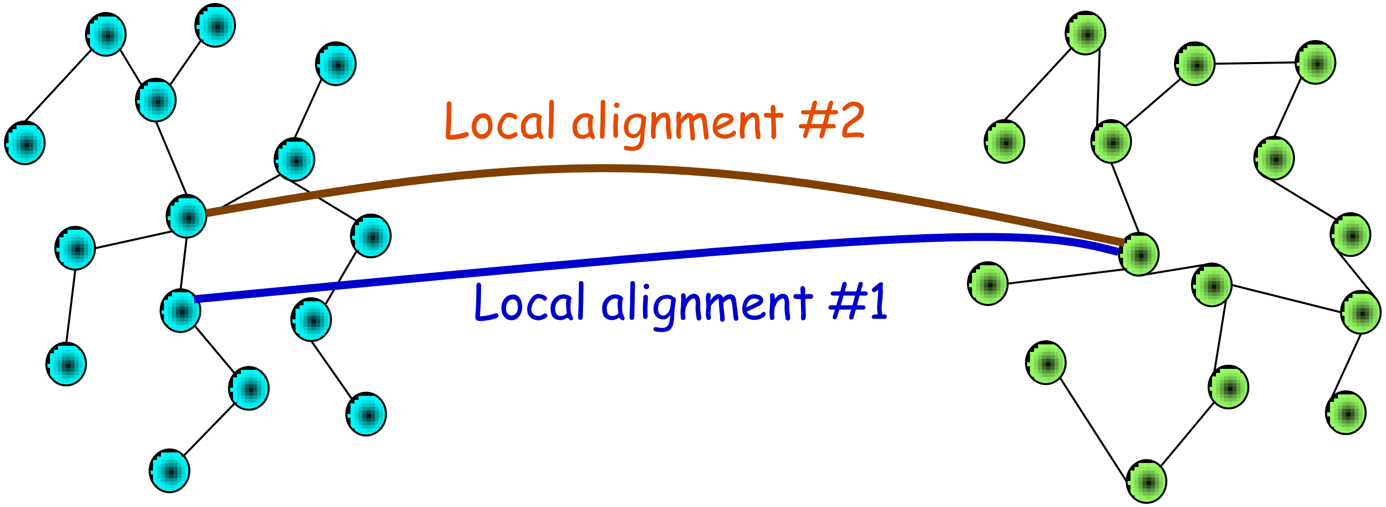
Local alignment #1

Local alignment #2

- Local vs. global alignment
  - Getting an overall match vs querying small patterns
- Parallels with sequence alignment (local vs. global)

IsoRank

Network Alignment: Local vs. Global

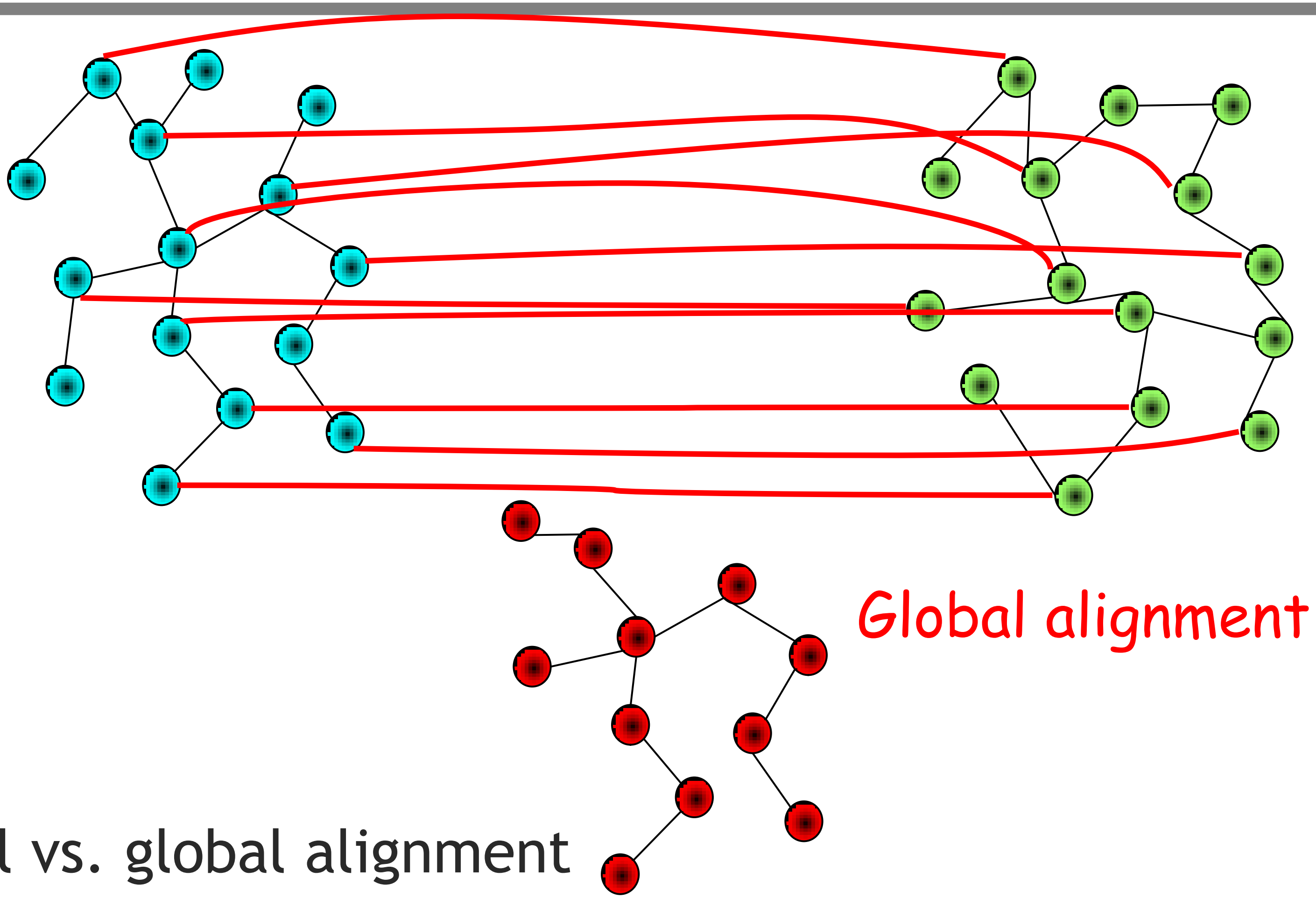


Local alignments: More than one mapping per node

- PathBlast (Kelley et al.)
- Koyuturk et al.
- Graemlin

- Local vs. global alignment
  - Getting an overall match vs qu
- Parallels with sequence alignment (local vs. global)

Network Alignment: Local vs. Global



- Local vs. global alignment
  - Getting an overall match vs querying small patterns
- Parallels with sequence alignment (local vs. global)

Problem Formulation

Given

1. Two or more undirected PPI graphs, one per species. Each graph contains all known PPIs for the species
2. [Optional] Pairwise similarity scores between proteins of the various species

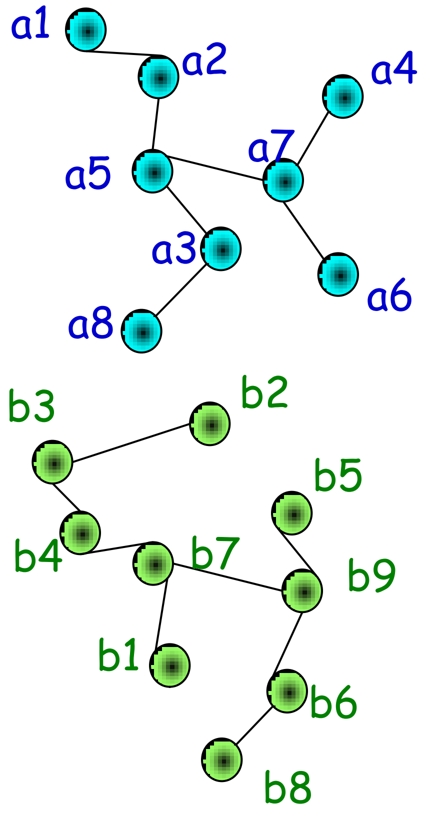
Find

1. Cross-species mapping between nodes of the various graphs. Must be closed under transitivity.
2. Estimate the common PPI subgraph across various species
3. [Optimality] Given just PPI graphs, maximize common subgraph size

Evaluation

1. Quality of mapping: 1) GO enrichment, 2) other orthologs
2. Coverage

Algorithm: IsoRank



Similarity Score

a5	b7	2.1
a5	b9	1.5
a3	b2	3.4

Stage 1

R

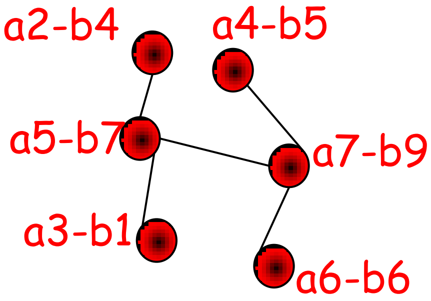
a5	b7	1e-2
a5	b1	2e-8
a5	b3	1e-7
a5	b9	1e-4
a3	b1	5e-4
a3	b6	3e-9
…		

Scores for each possible node mapping

Stage 2

a5	b7
a3	b1
a7	b9
a6	b6
a4	b5
a2	b4

mapping

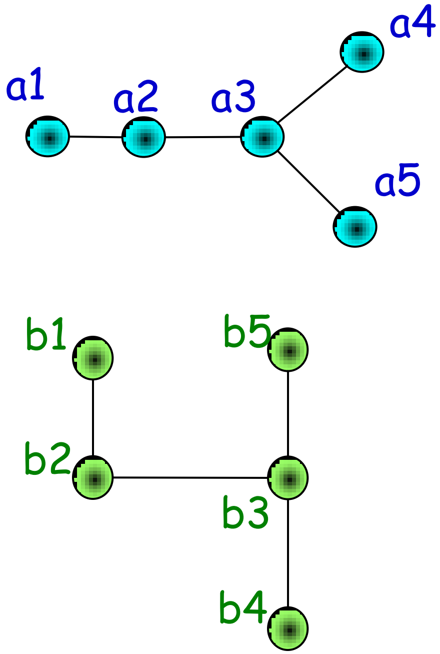


Computing R: just network similarity

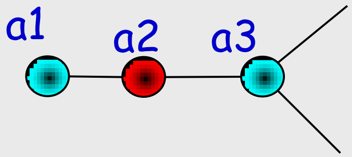
- R_{ij} depends on neighborhoods of i and j

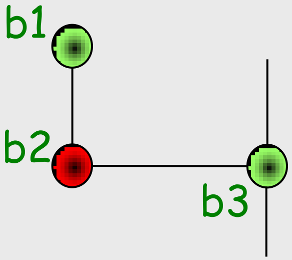
$$R_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{|N(u)||N(v)|} R_{uv}$$

- N(a) is the set of neighbors of a

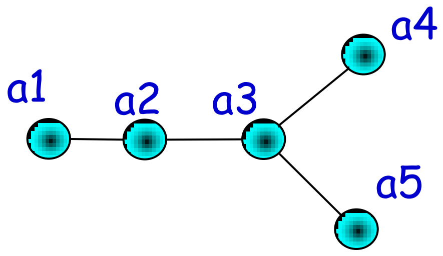


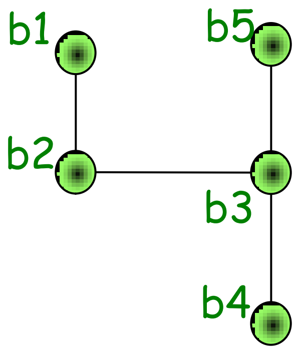
$$R_{a2,b2} = \frac{1}{1\times1} R_{a1,b1} + \frac{1}{1\times3} R_{a1,b3}$$
$$+ \frac{1}{3\times1} R_{a3,b1} + \frac{1}{3\times3} R_{a3,b3}$$





Example: Computed R_{ij} values



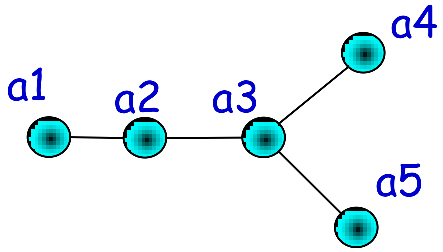


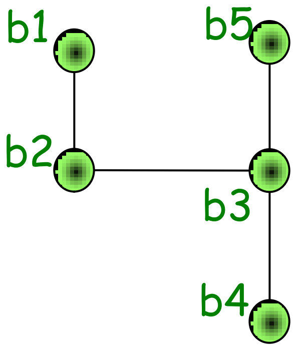
R

	b1	b2	b3	b4	b5
a1	0.0312		0.0937		
a2		0.1250		0.0625	0.0625
a3	0.0937		0.2813		
a4		0.0625		0.0312	0.0312
a5		0.0625		0.0312	0.0312

Empty cell indicates $R_{ij} = 0$

Example: Computed R_{ij} values





R

	b1	b2	b3	b4	b5
a1	0.0312		0.0937		
a2		0.1250		0.0625	0.0625
a3	0.0937		0.2813		
a4		0.0625		0.0312	0.0312
a5		0.0625		0.0312	0.0312

Empty cell indicates $R_{ij} = 0$

Computing R is an eigenvalue problem

- The equations for R describe an eigenvalue problem

$$R = AR$$

$$A[ij][uv] = \frac{1}{|N(u)||N(v)|}$$

$$size(A) = N_1 N_2 \times N_1 N_2$$

N1 = # nodes in Graph 1
N2 = # nodes in Graph 2

- A is about 10^8x10^8 when aligning yeast and fly networks
  - However, both A and R are very sparse.
  - We use the Power method to efficiently compute R
- Extension to weighted edges is straightforward

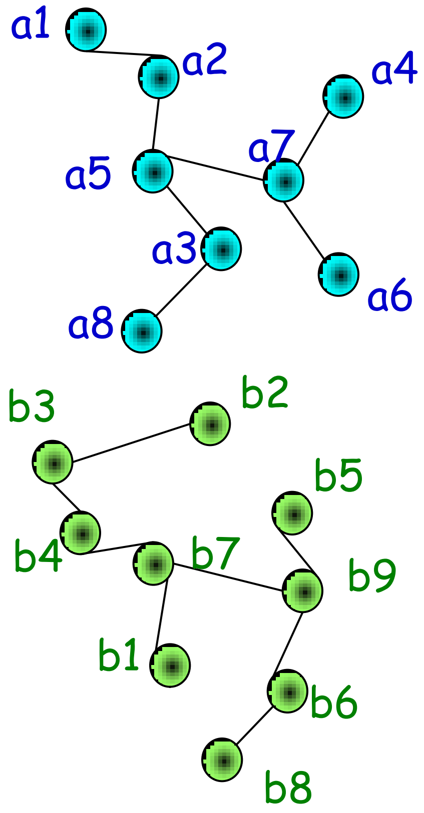
Computing R: including sequence data

- Let B_{ij} = similarity score between i (from graph #1) and j from (graph #2)
- $E_{ij} = B_{ij} / |B|$

$$R = \alpha AR + (1 - \alpha) E$$

$$0 \leq \alpha \leq 1$$

Algorithm: IsoRank



Similarity Score

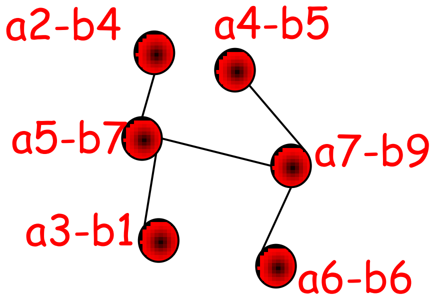
a5	b7	2.1
a5	b9	1.5
a3	b2	3.4

a5	b7	1e-2
a5	b1	2e-8
a5	b3	1e-7
a5	b9	1e-4
a3	b1	5e-4
a3	b6	3e-9
...		

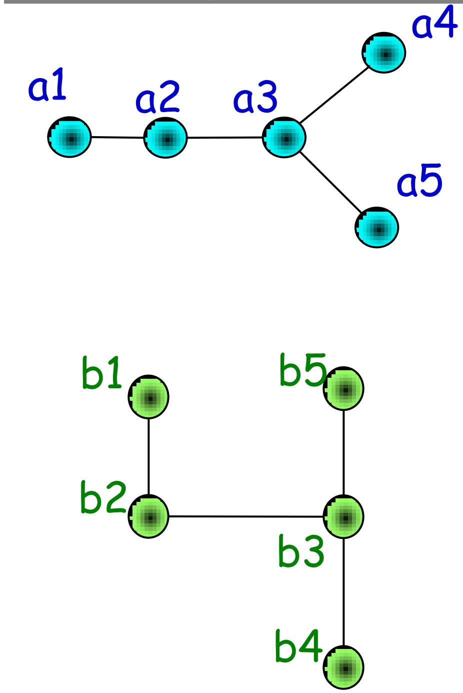
Scores for each possible node mapping

a5	b7
a3	b1
a7	b9
a6	b6
a4	b5
a2	b4

mapping

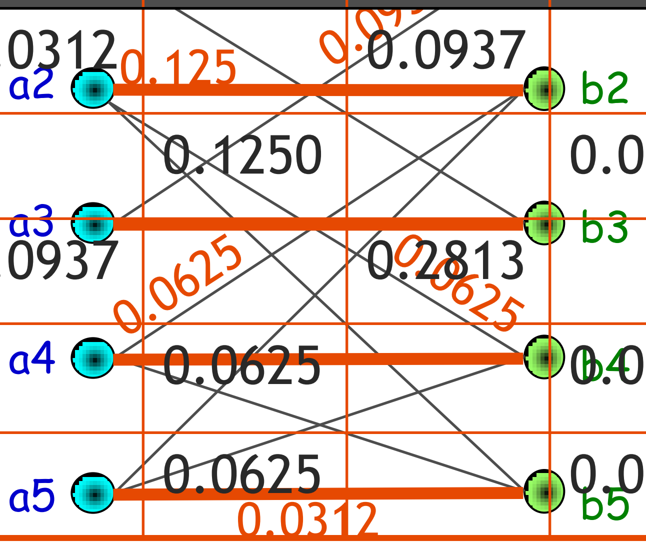


Stage 2: Two-species case
Compute one-to-one mapping



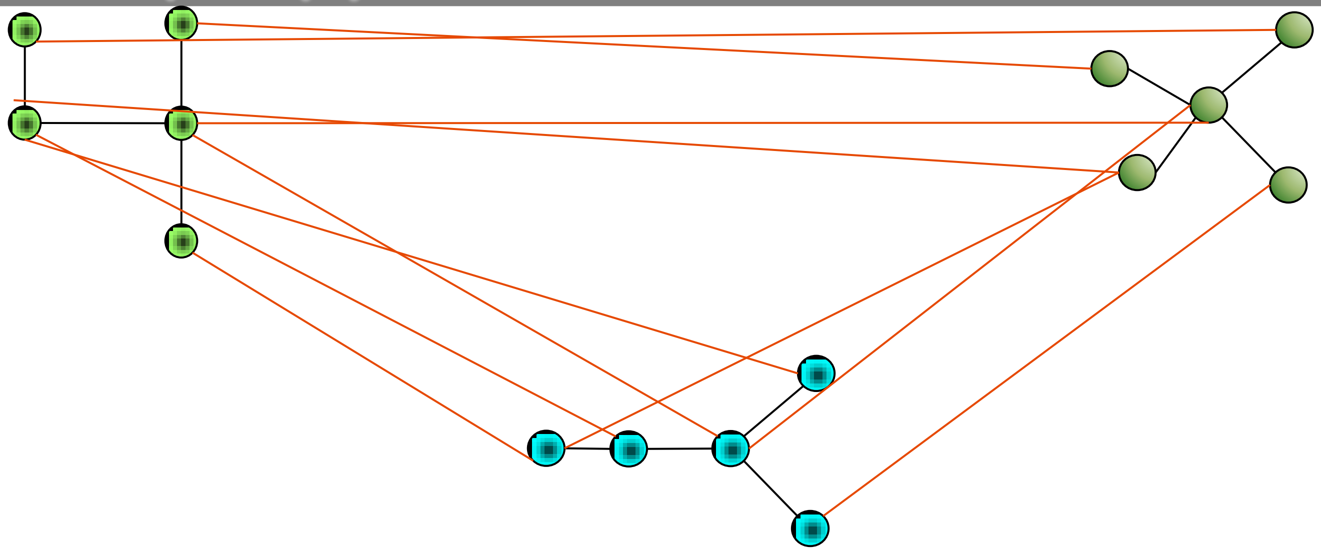
R

	b1	b2	b3	b4	b5
a1	0.0312		0.0937		
a2		0.1250		0.0625	0.0625
a3	0.0937		0.2813		
a4		0.0625		0.0312	0.0312
a5		0.0625		0.0312	0.0312



- Strategy #1: Max Weighted Bipartite matching
- Strategy #2: Greedy
  - At each iteration, pick the highest weight edge between nodes not yet picked

Stage 2: Multiple species case: Greedy approach



- From the k-partite graph described by R,
  - Pick largest weight edge R_{ij}
  - In every other species, find if a node is the best match to both *i* and *j*. If such a node exists, add it.
  - Add secondary nodes which have good-enough matches to selected nodes

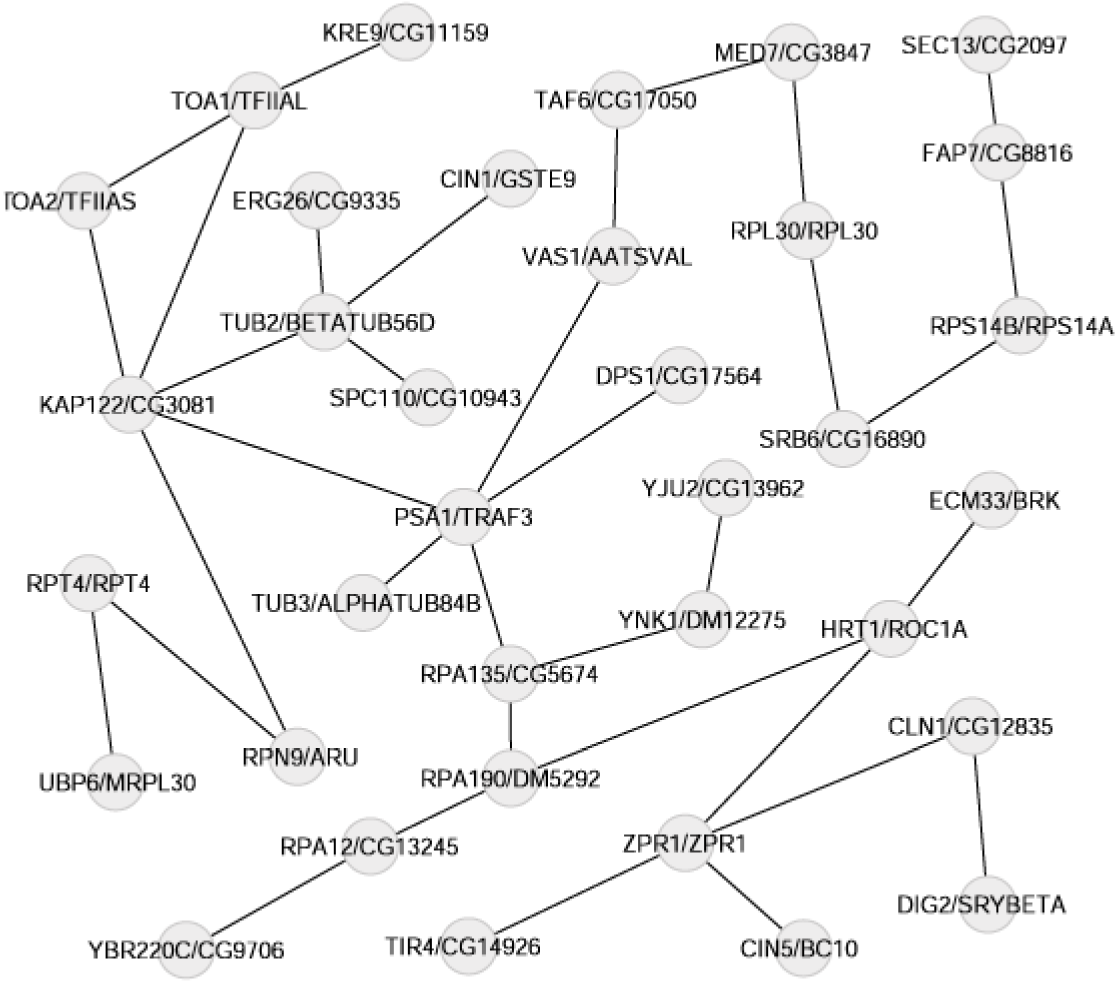
Stage 2:Multiple species case: IsoRankN

Find high-weight near-cliques using spectral technique:

- For each node *v*, construct its Star S_v, consisting of nodes with largest-weight edges to it
- At each step:
    - Pick the star S_v with highest total weight
    - Spectral partitioning to identify approx-clique S^*_v that contains *v*
        - Use Personalized PageRank algorithm
    - Join two sets $S^*_{v\,1}$ *and* $S^*_{v\,2}$ if their nodes have large-weight edges to each other

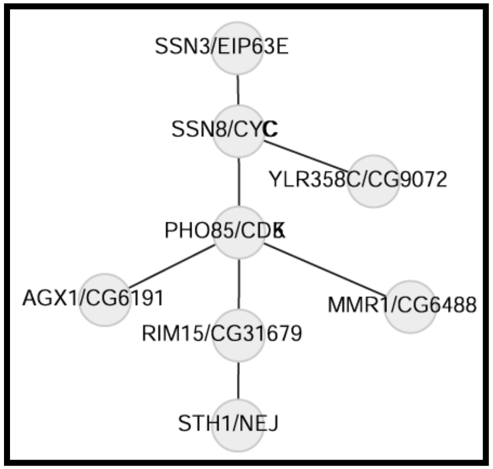
Results: 2-species case: Yeast-Fly alignment

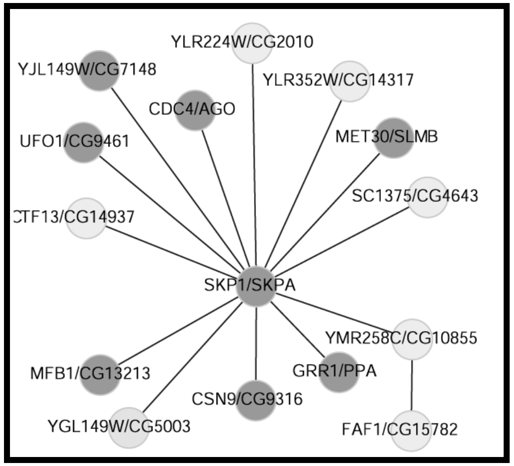
- # of edges in the common subgraph: 1420

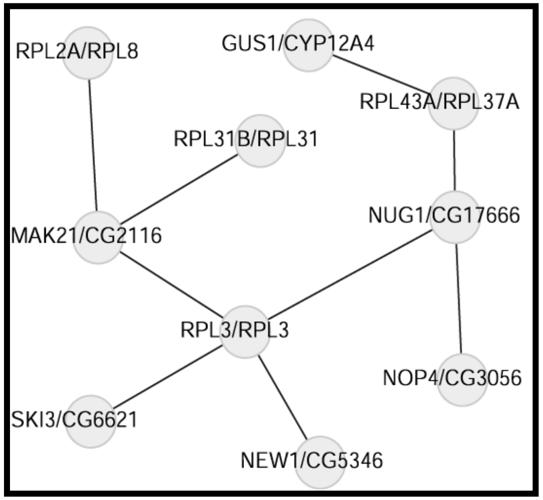


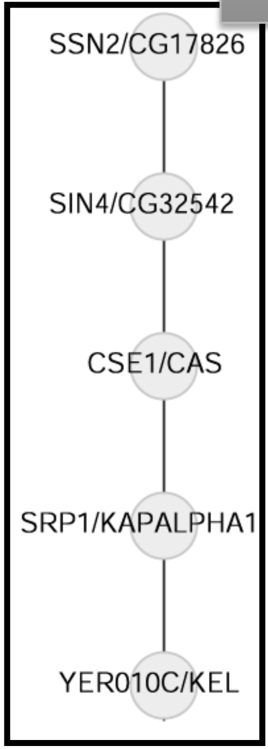
Largest connected component in Yeast/Fly alignment

Various Topologies Are Found









Existing local alignment methods often find only specific topologies

IsoRankN: functional coherence

$$H(S_v^*) = -\sum_t p_t \log p_t$$

where p_t is the fraction of times GO/KEGG term t occurs in node-set

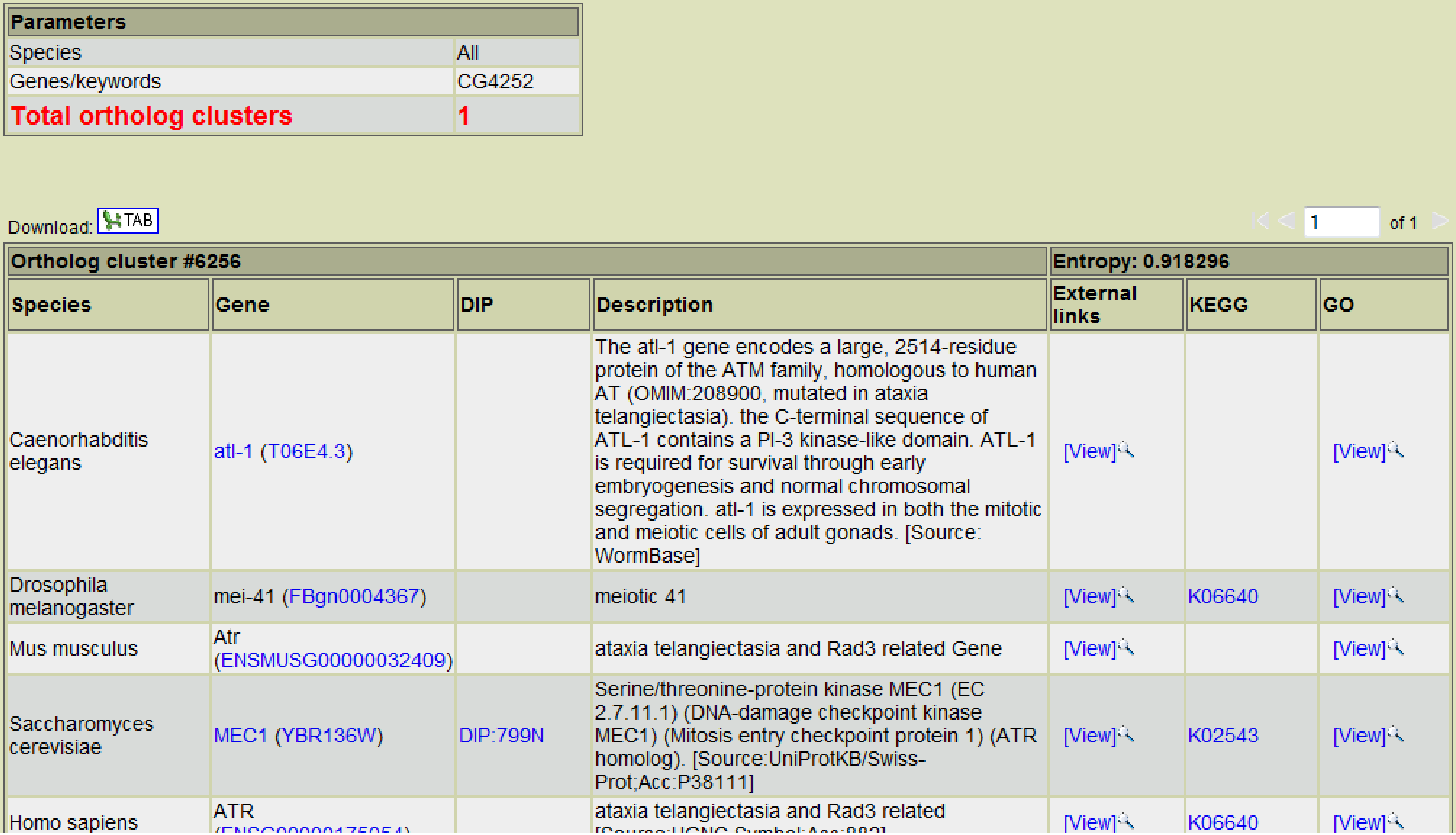
	IsoRankN	IsoRank	Graemlin-1K	Graemlin-2K	NetworkBLAST-M
Normalized GO/KEGG entropy	**0.179**	0.359	0.451	0.357	0.554
Exact Cluster Ratio	**0.380**	0.253	0.306	0.355	0.291

IsoRankN: coverage

k	IsoRankN	IsoRank	Graemlin-1K	Graemlin-2K
2	8739	**20580**	4650	5899
3	**13533**	13391	5414	5072
4	**13991**	15422	5371	2067
5	**12715**	9744	1467	78
Total	48978	59539	20903	16026

Number of proteins in clusters with exactly *k* species

IsoBase

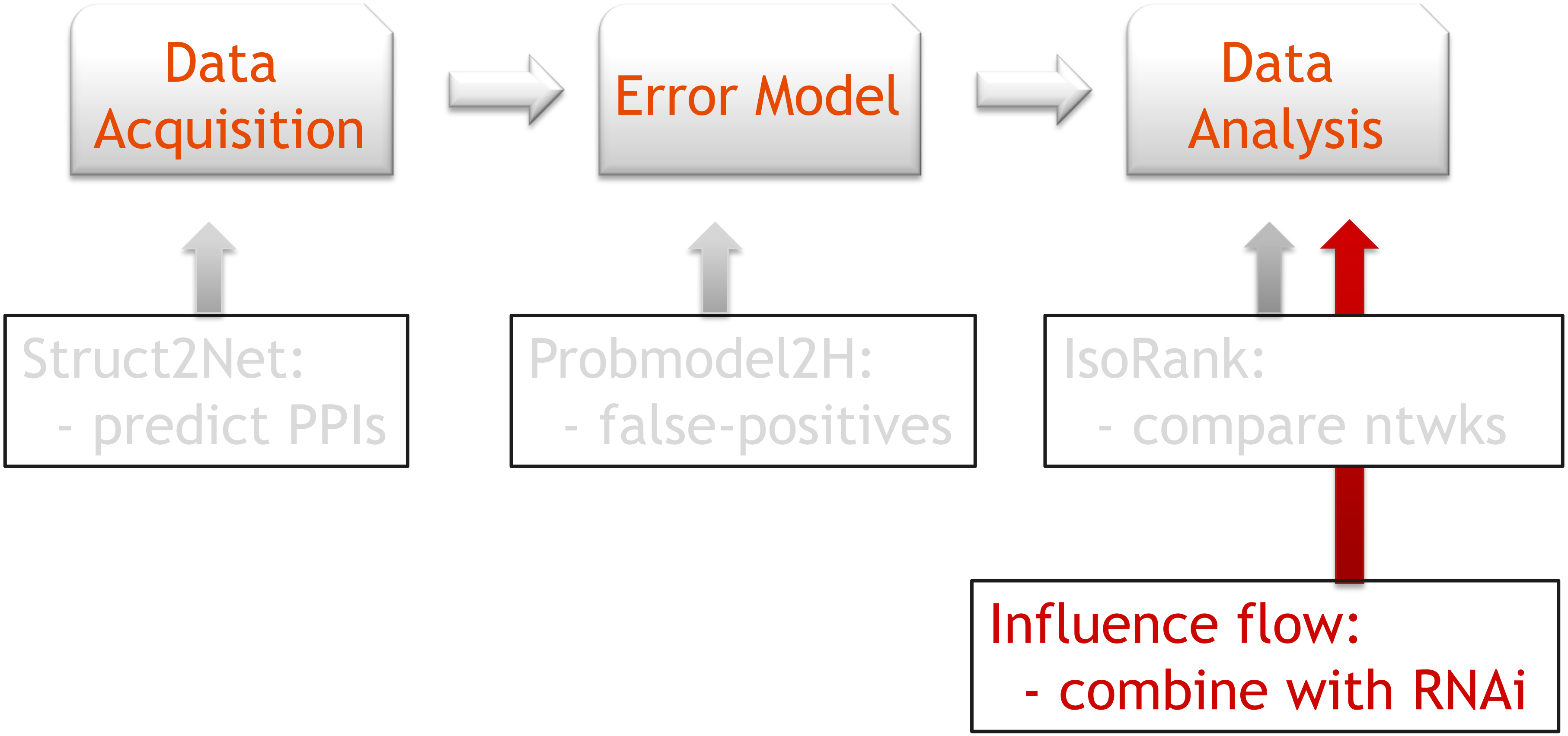


Parameters	
Species	All
Genes/keywords	CG4252
Total ortholog clusters	**1**

Ortholog cluster #6256				Entropy: 0.918296		
Species	**Gene**	**DIP**	**Description**	**External links**	**KEGG**	**GO**
Caenorhabditis elegans	atl-1 (T06E4.3)		The atl-1 gene encodes a large, 2514-residue protein of the ATM family, homologous to human AT (OMIM:208900, mutated in ataxia telangiectasia). the C-terminal sequence of ATL-1 contains a PI-3 kinase-like domain. ATL-1 is required for survival through early embryogenesis and normal chromosomal segregation. atl-1 is expressed in both the mitotic and meiotic cells of adult gonads. [Source: WormBase]	[View]		[View]
Drosophila melanogaster	mei-41 (FBgn0004367)		meiotic 41	[View]	K06640	[View]
Mus musculus	Atr (ENSMUSG00000032409)		ataxia telangiectasia and Rad3 related Gene	[View]		[View]
Saccharomyces cerevisiae	MEC1 (YBR136W)	DIP:799N	Serine/threonine-protein kinase MEC1 (EC 2.7.11.1) (DNA-damage checkpoint kinase MEC1) (Mitosis entry checkpoint protein 1) (ATR homolog). [Source:UniProtKB/Swiss-Prot;Acc:P38111]	[View]	K02543	[View]
Homo sapiens	ATR (ENSG00000175054)		ataxia telangiectasia and Rad3 related [Source:HGNC Symbol;Acc:882]	[View]	K06640	[View]

Outline

- Introduction to Protein Interactions
- Algorithms for PPI Networks:



Data Acquisition

Error Model

Data Analysis

Struct2Net:
- predict PPIs

Probmodel2H:
- false-positives

IsoRank:
- compare ntwks

Influence flow:
- combine with RNAi

Influence Flow

Goal: generate hypotheses about signaling networks' structure

Why?

- Understanding signaling networks is very valuable
- Old view of signaling cascade seems too naïve, need a network picture

How?

- RNA interference data provides signaling information
- PPI provides routing information
- Look for a simple explanation that is consistent with both

Acknowledgments

Infl. Flow

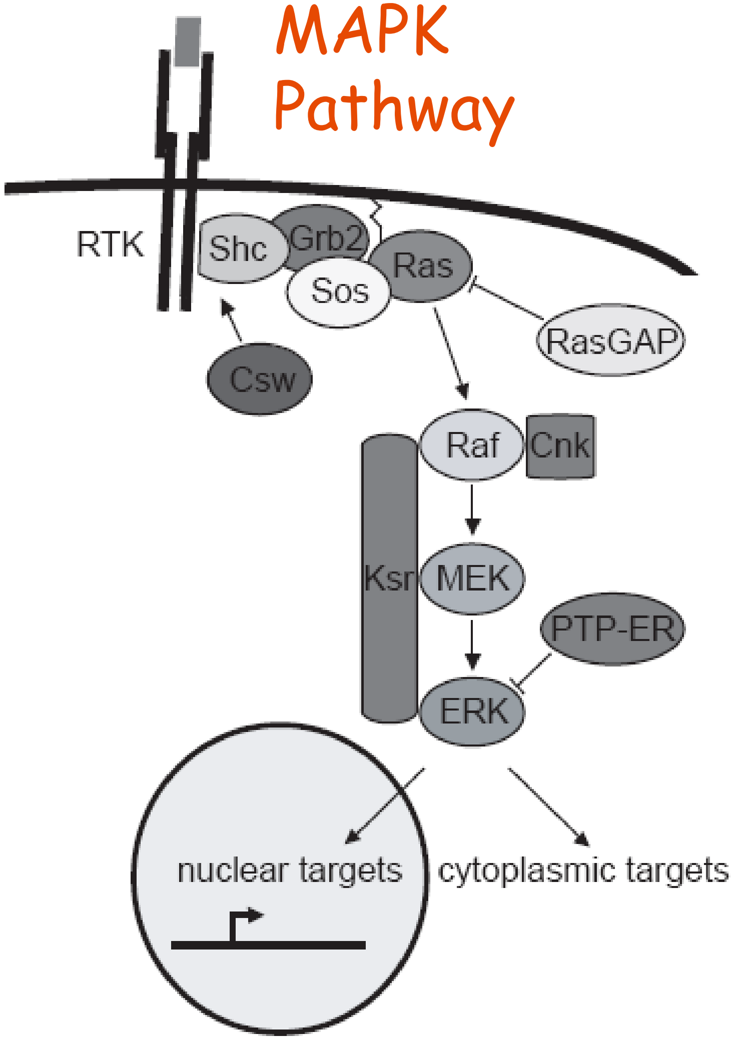
- Collaborators:
  - Adam Friedman, Norbert Perrimon & Bonnie Berger
  - Future Work in collaboration with George Tucker and Vinu Arunachalam
- Previously presented/published:
  - ISMB 2007 (highlights track)

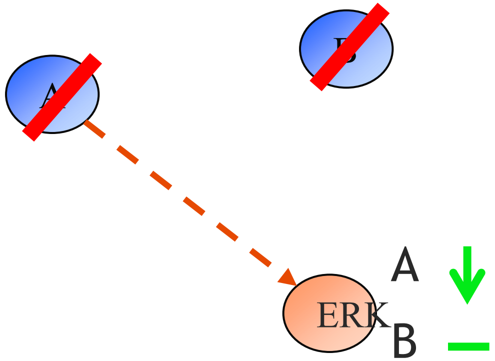
Other work:

- Yeang et al. (2004)
- Ourfali et al (2007)
- Yeger-Lotel et al. (2009)

Screening for MAPK pathway regulators with RNAi

MAPK Pathway





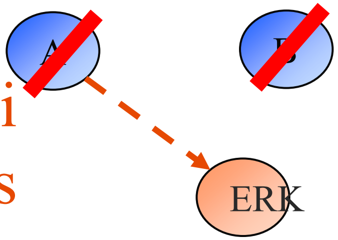
A	3.1
D	2.0
…	…

Whole genome screen for regulators of MAPK pathway

- hundreds of hits (331)
- 56% of genes have unknown function

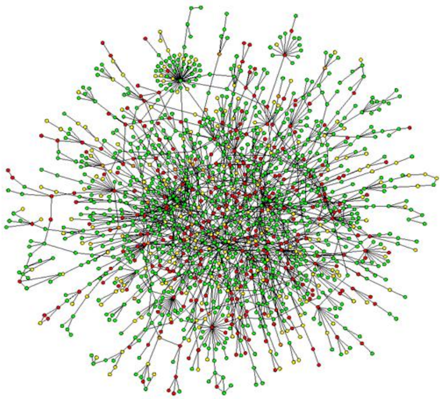
Goal: a simple explanation consistent with data and known biology

RNAi hits

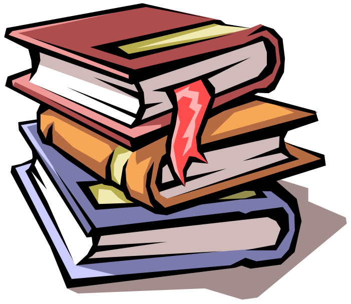


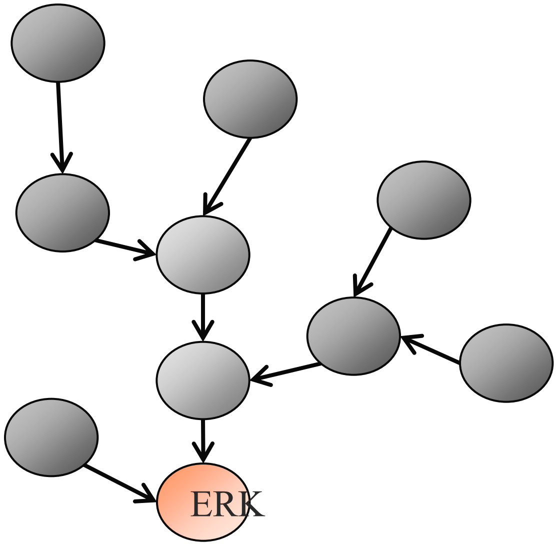
A	3.1
D	2.0
…	…

PPI network



Biological info





Influence Network

Problem Formulation

Probmodel 2H

Given

1. Undirected PPI data for the species
    1. [Optional] Augment with cross-species PPI data or expression data
2. The end-effector G_p of the pathway ***P*** being investigated
3. RNAi scores, with score S_i indicates impact of knocking-down gene G_i on the activity of the end-effector G_p
4. Known, high-confidence estimate of ***P***'s core cascade

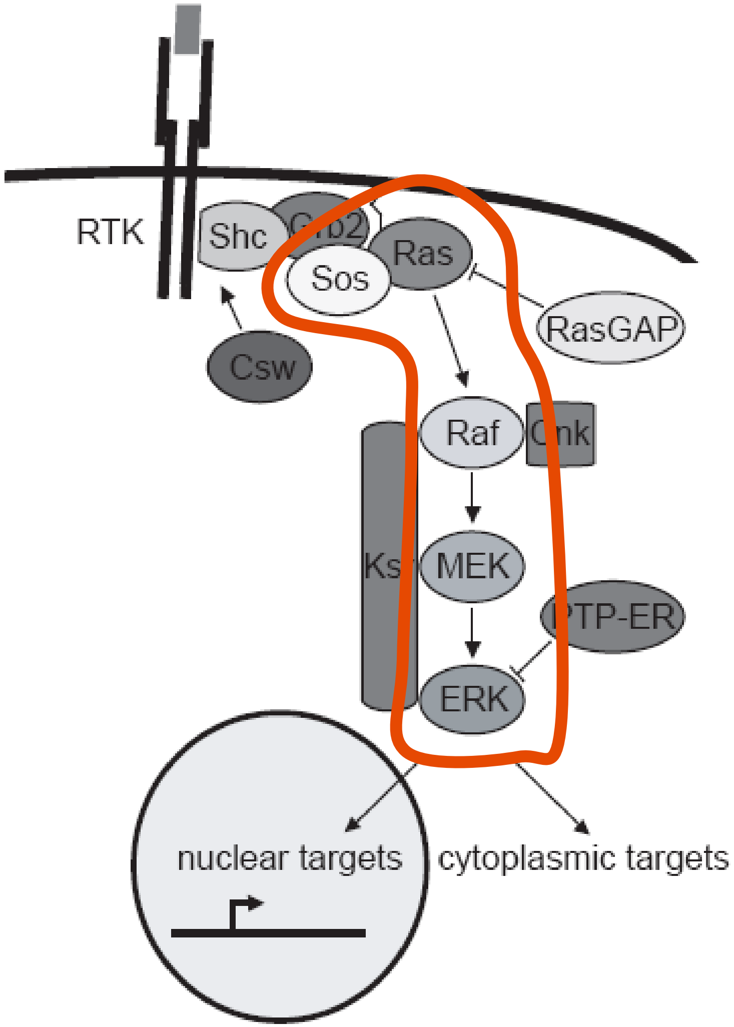
Find

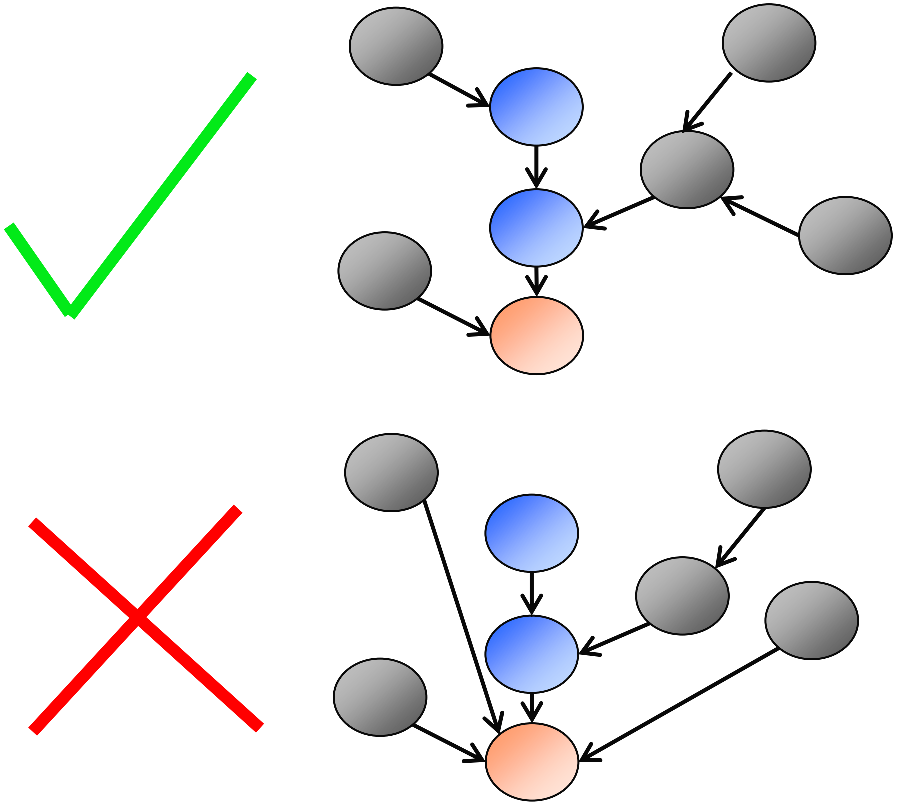
1. A directed, sparse network with edges directed along the way signal might flow, finally ending in the end-effector G_p

Evaluation

1. Provide only a subset of the pathway's known components as input. See if the remaining components are discovered

Using The Core Cascade

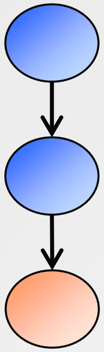




Core cascade should be the central trunk of the influence network

Algorithm: Preliminary Processing

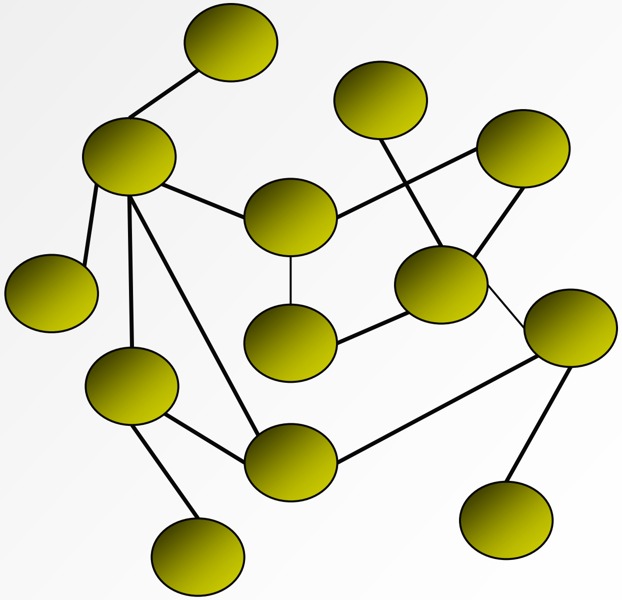
Core cascade

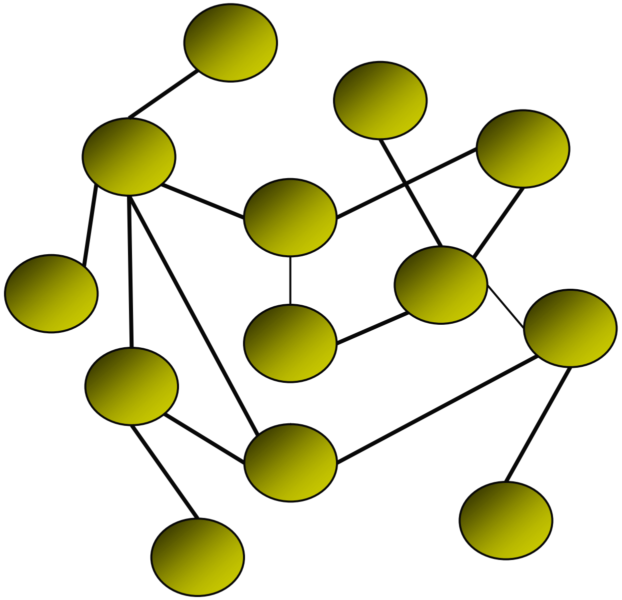


RNAi hits

A	3.1
D	2.0
…	…

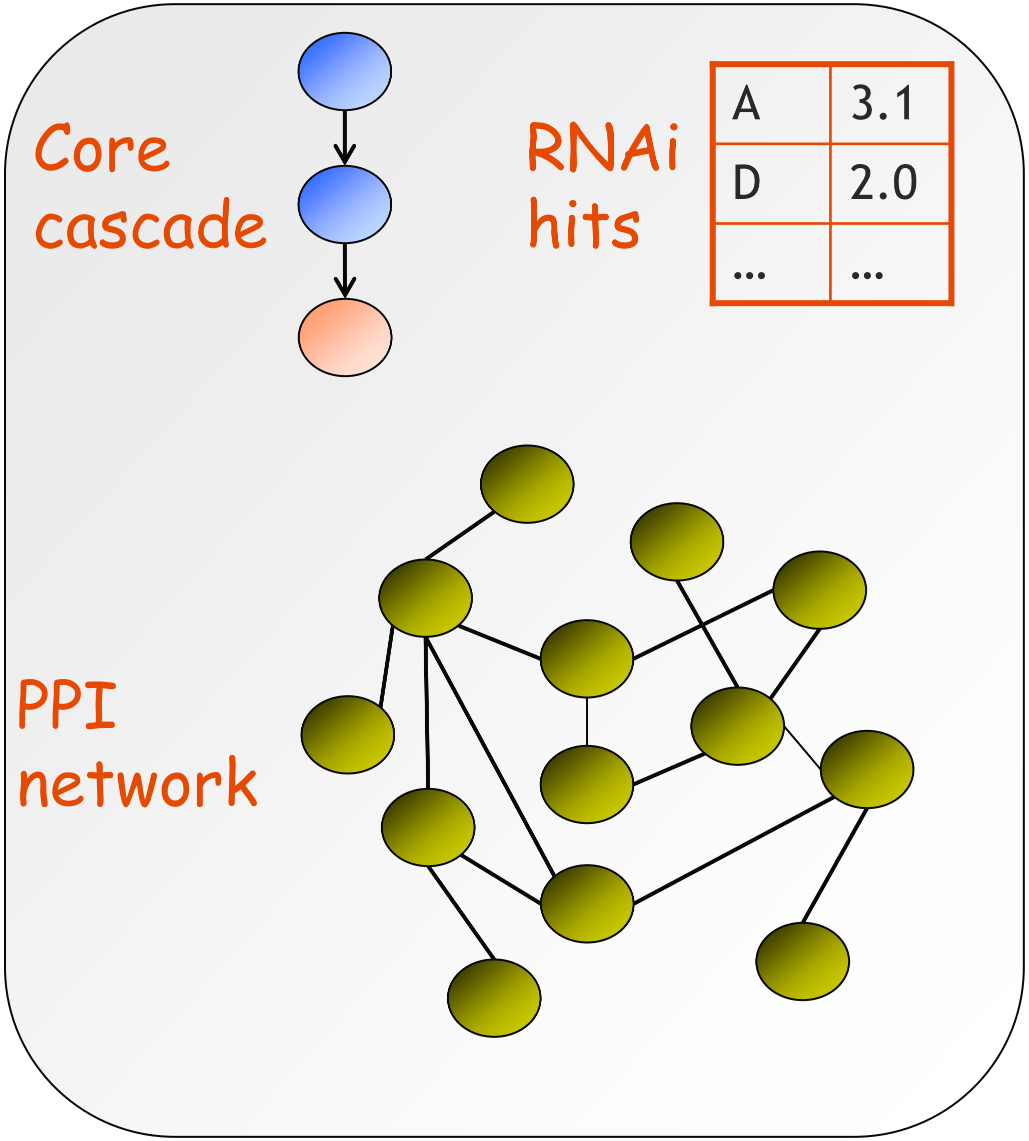
PPI network



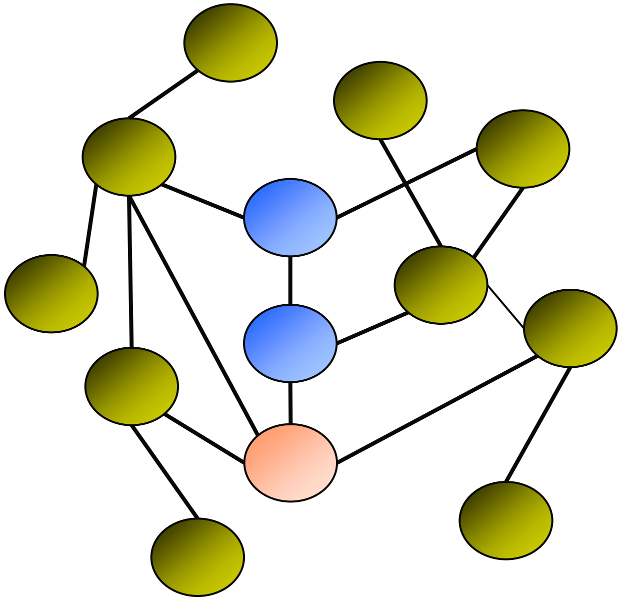


Occam's Razor:
simple, sparse solution

Algorithm: Preliminary Processing



A	3.1
D	2.0
…	…

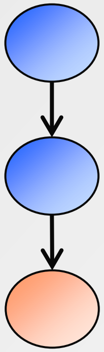


Add core cascade

**Occam's Razor:
simple, sparse solution**

Algorithm: Preliminary Processing

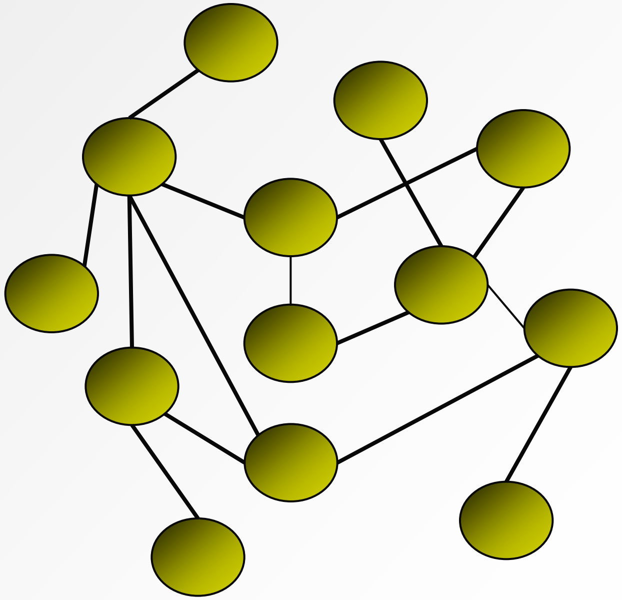
Core cascade

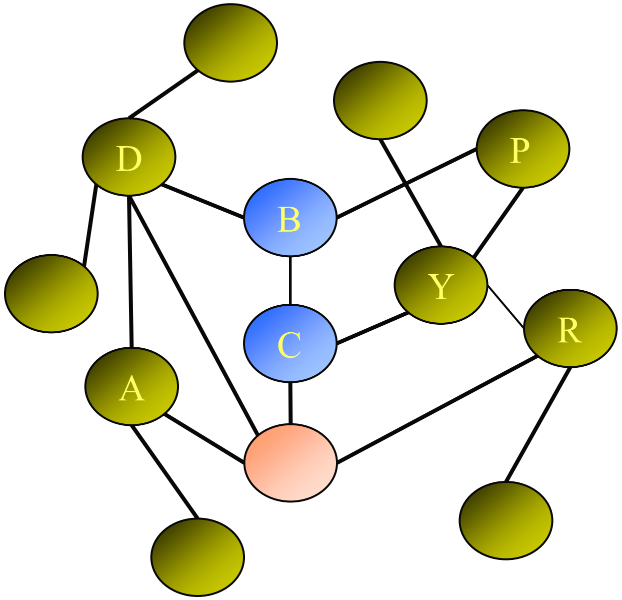


RNAi hits

A	3.1
D	2.0
…	…

PPI network



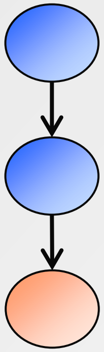


Map RNAi data

Occam's Razor: simple, sparse solution

Algorithm: Preliminary Processing

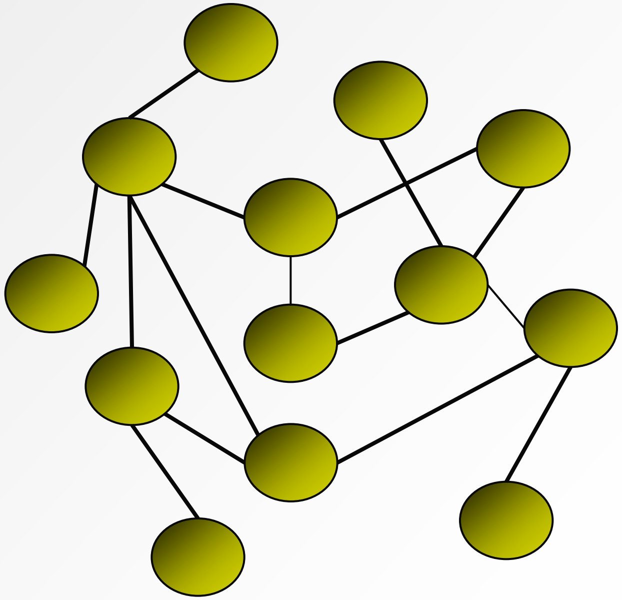
Core cascade

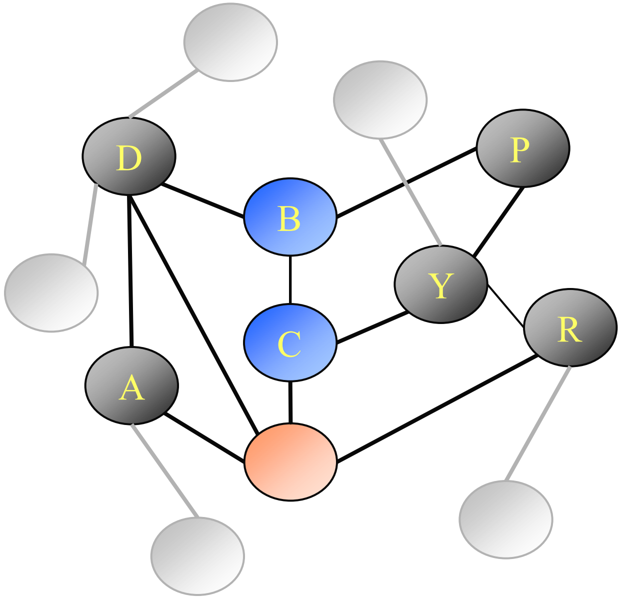


RNAi hits

A	3.1
D	2.0
…	…

PPI network

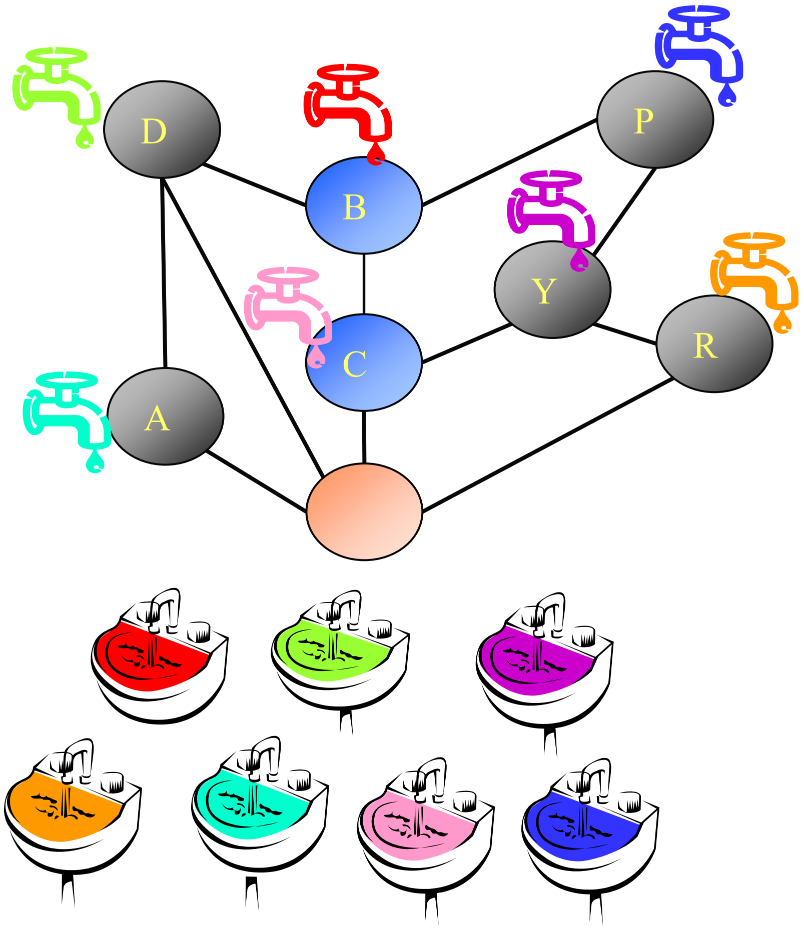


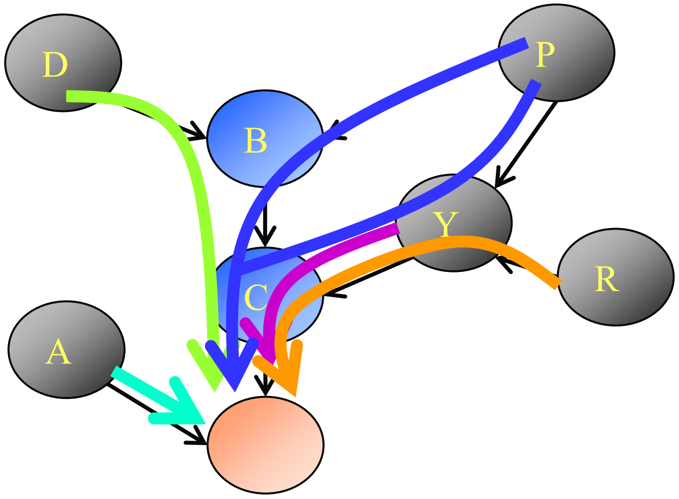


Select RNAi subgraph

Occam's Razor:
simple, sparse solution

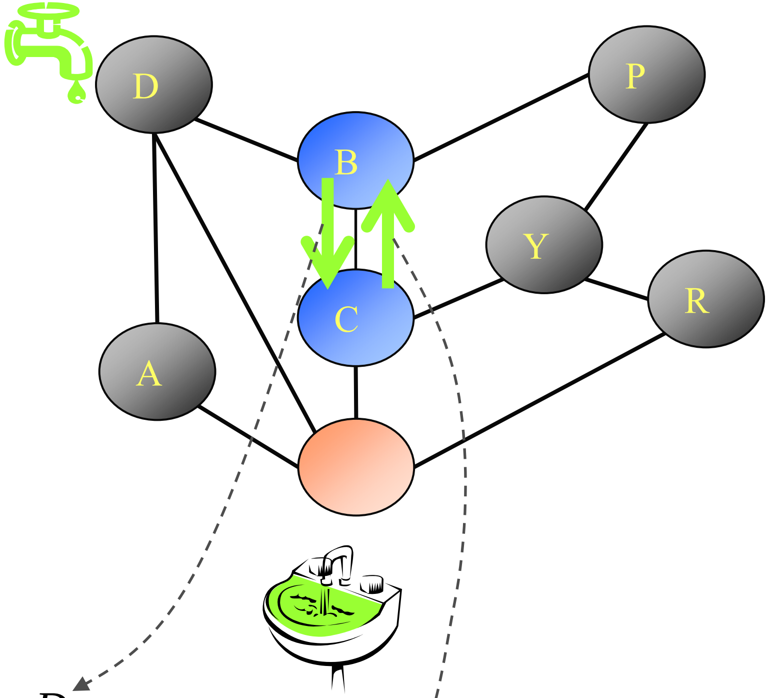
Influence Flow: prune edges and assign direction





Multi-commodity flow

Integer Linear Program



from source $$\sum_{e\in\delta^-(r)} f_e^k - \sum_{e\in\delta^+(r)} f_e^k = 1$$

conservation $$\sum_{e\in\delta^-(v)} f_e^k - \sum_{e\in\delta^+(v)} f_e^k = 0$$

into sink $$\sum_{e\in\delta^-(k)} f_e^k - \sum_{e\in\delta^+(k)} f_e^k = -1$$

capacity $$f_{ij}^k \le y_{ij}$$

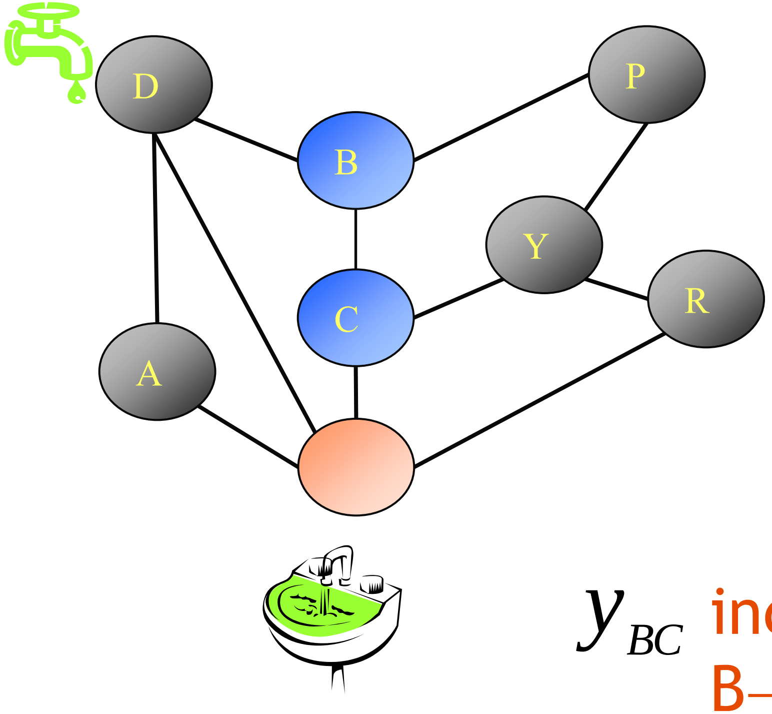
$$y_e = 0,1$$

f_{BC}^D = flow of type D, along B→C

f_{CB}^D = flow of type D, along C→B

y_{BC} indicates if edge B-C with direction B→C is selected

Look for as few edges as possible



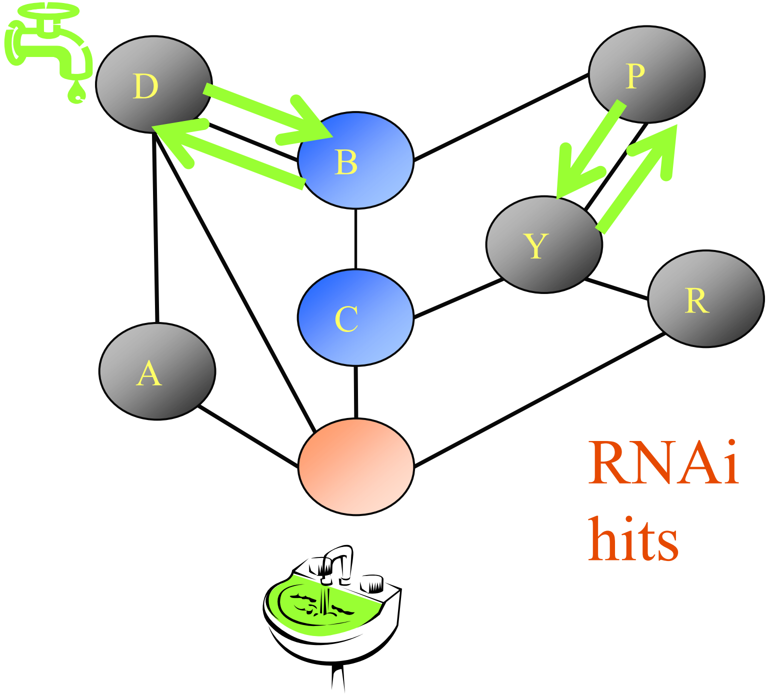
previous constraints
&

n-1 edges
⇒ tree

$$\sum y_e = n - 1$$

y_{BC} indicates if edge B-C with direction B→C is selected

Imposing directionality using RNAi Scores



RNAi hits

P	3.5
Y	2.0
…	…

f_{PY}^{D} = flow of type D, along P→Y

previous constraints &

$$f_{ij}^{k} = 0 \ \forall k$$

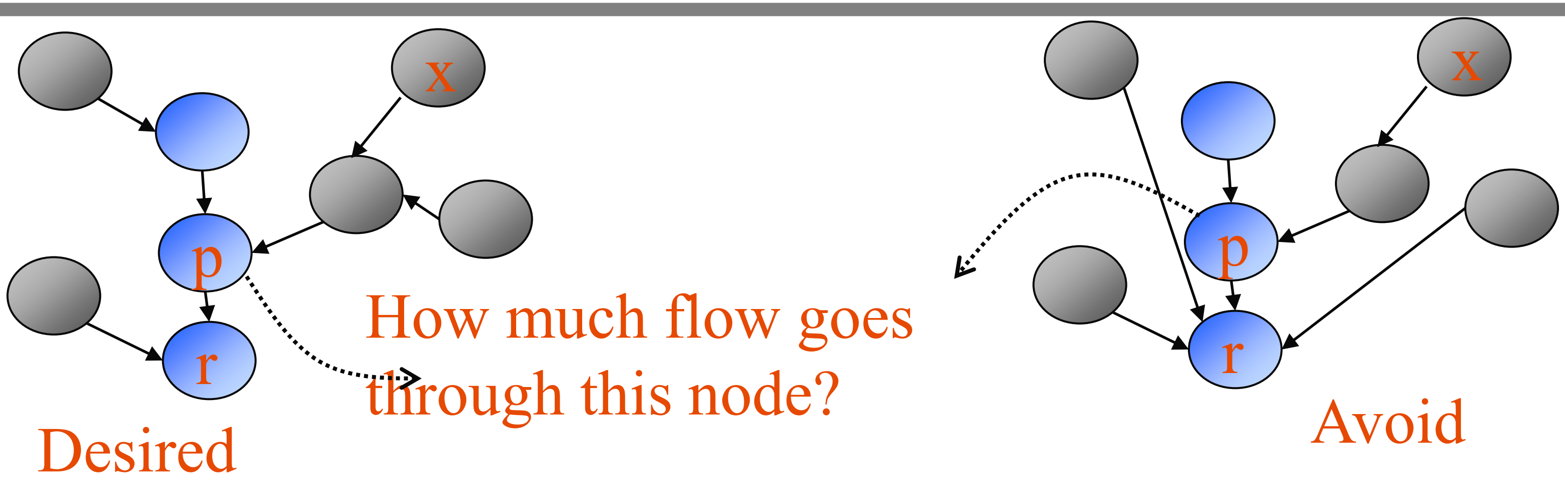
if

$$s_i - s_j < \Delta$$

or

i in core cascade

Connections to the core cascade



How much flow goes through this node?

Desired

Avoid

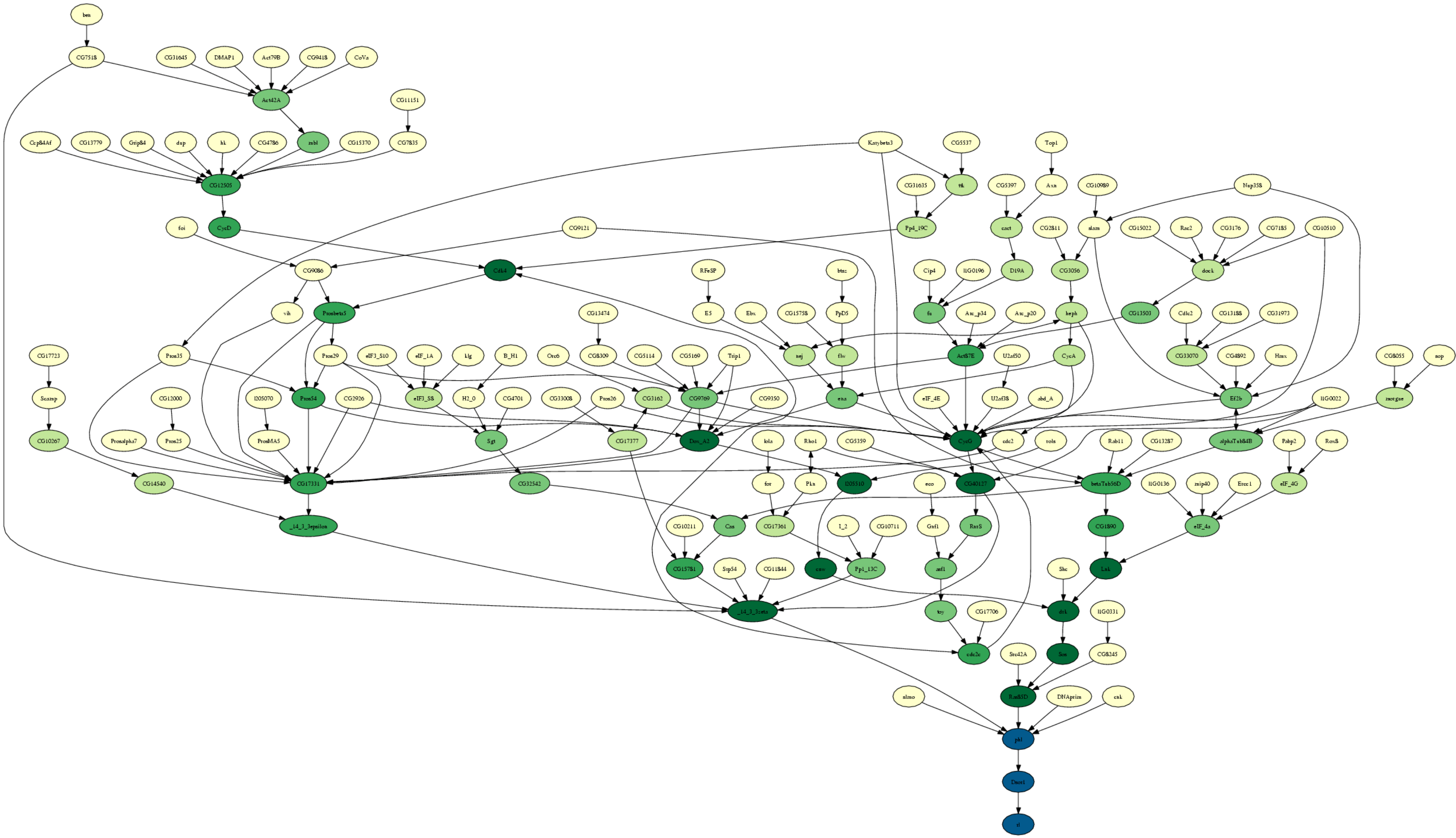
$$\sum_{e\in\delta^-(k)} f_e^k = z_k$$

for all x not in core cascade

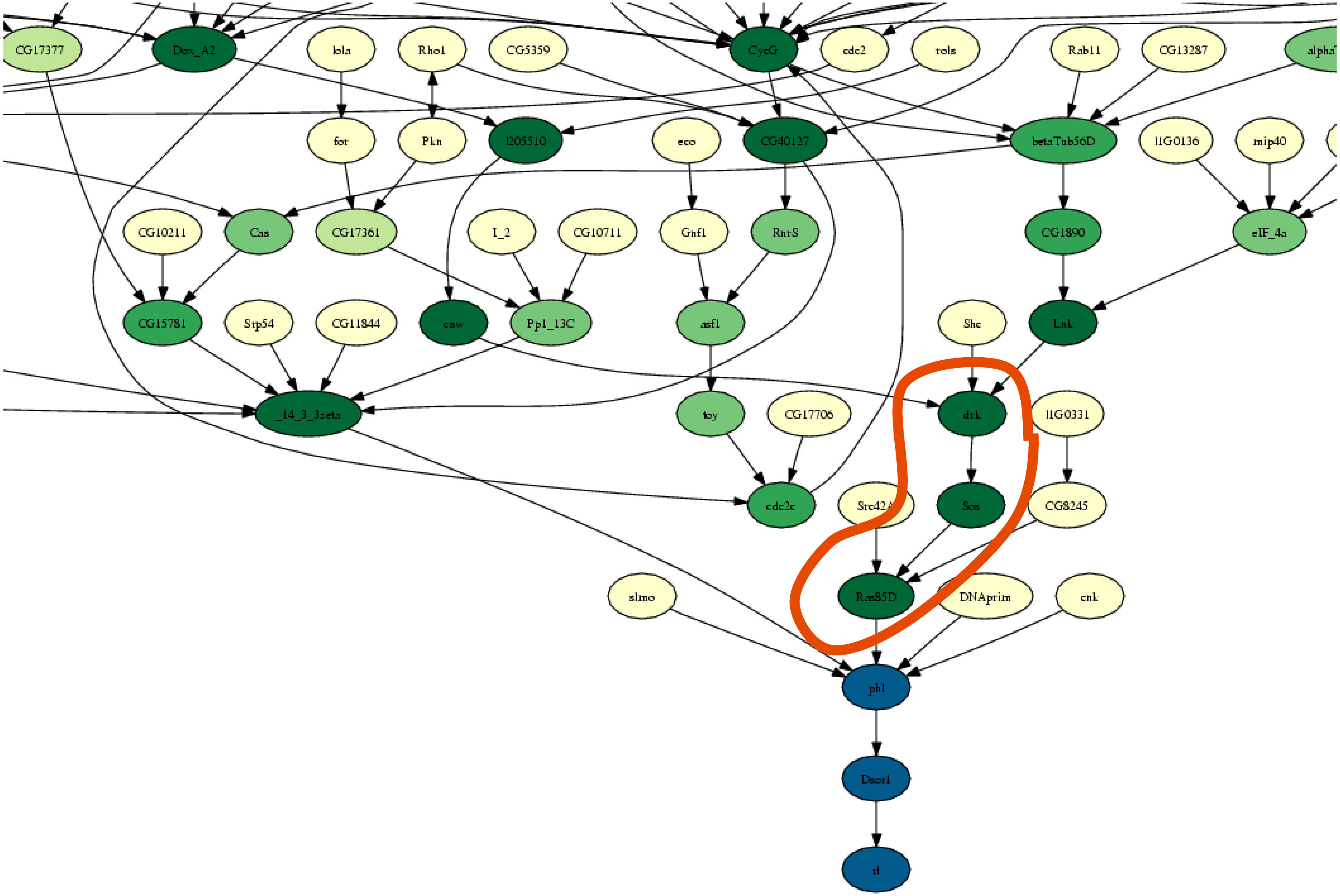
$$z_p - z_x \geq h$$

Maximize h

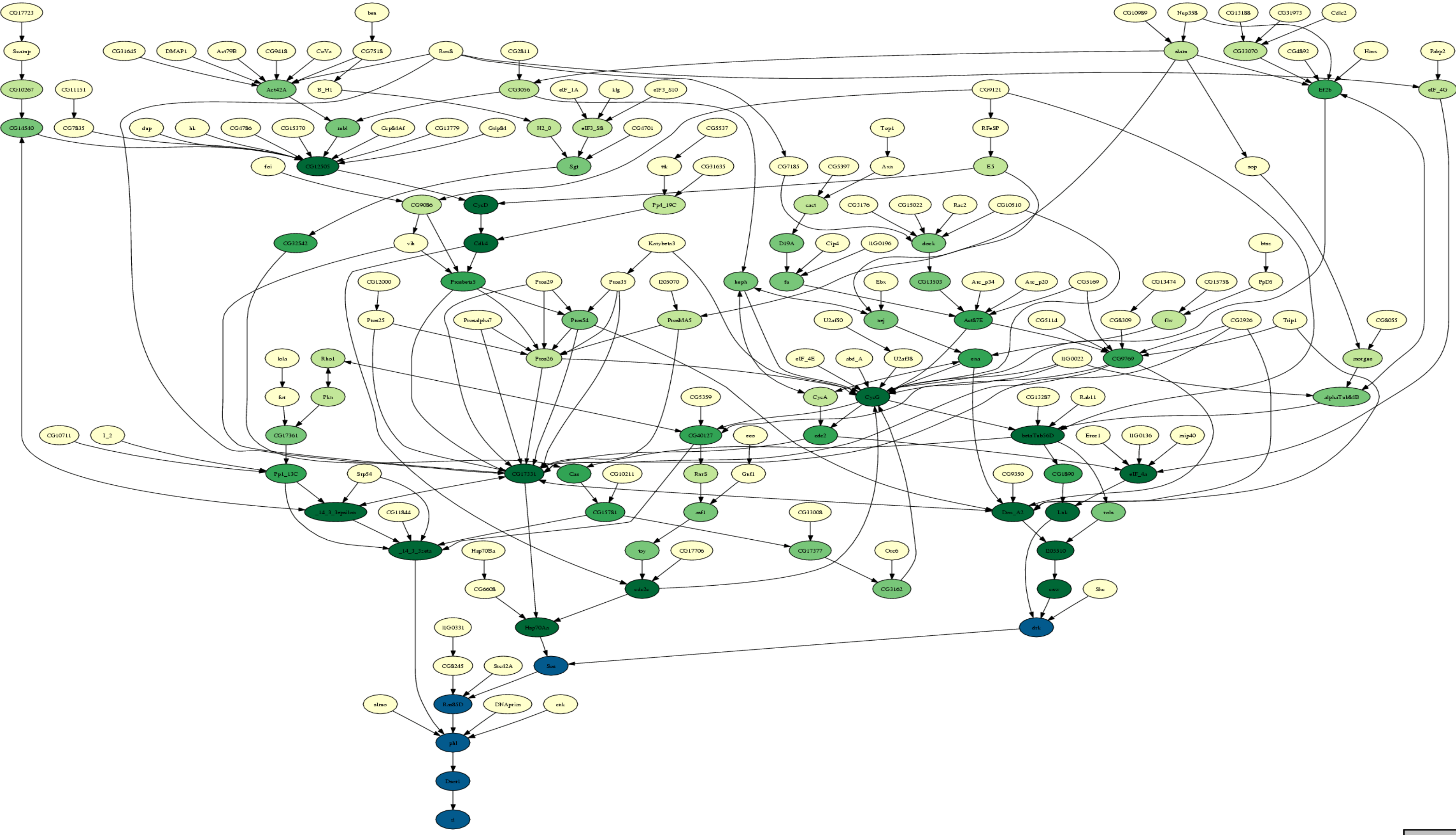
Results: can rediscover parts of the core cascade



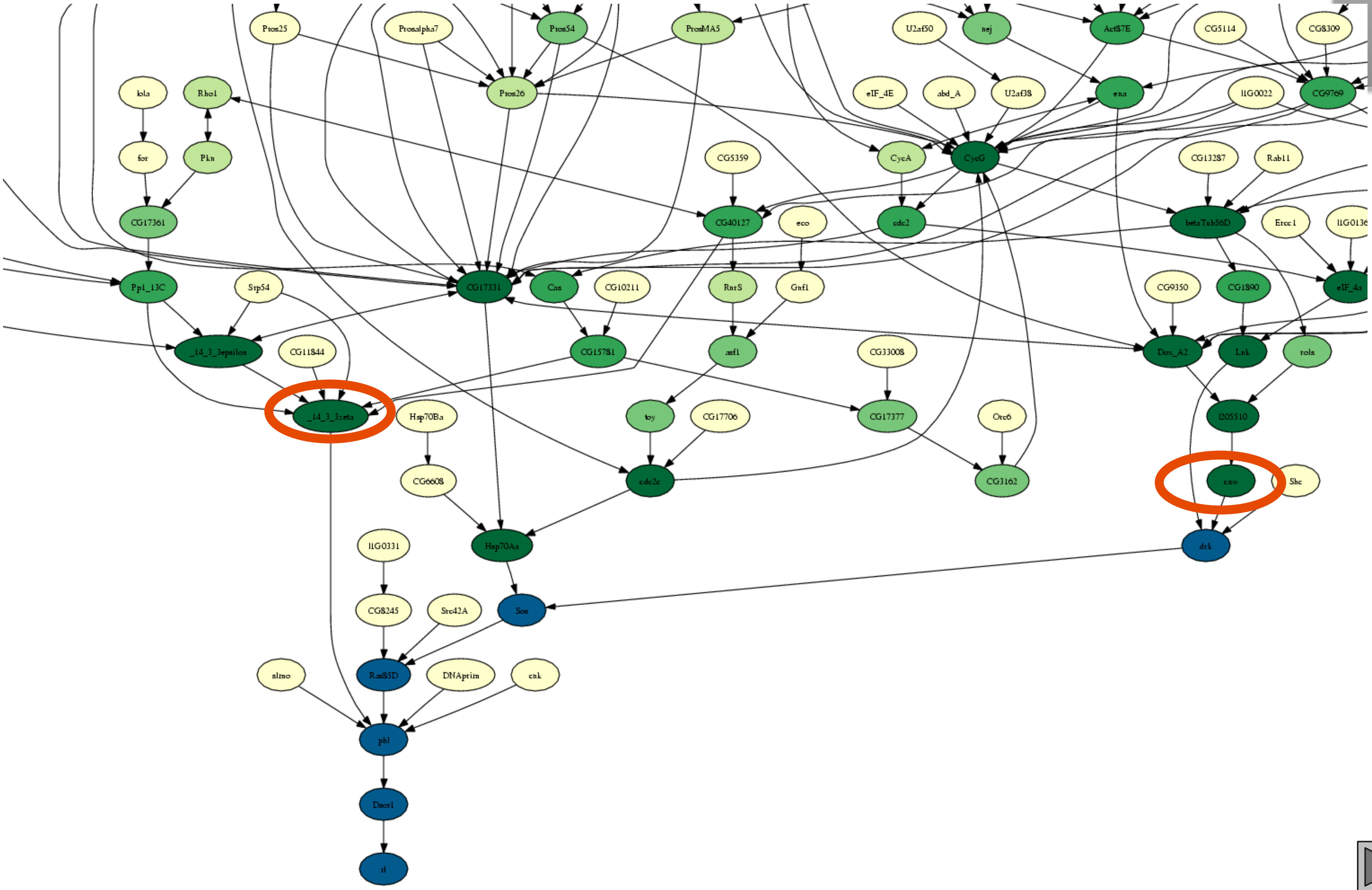
Results: can rediscover parts of the core cascade



Results: using full MAPK cascade

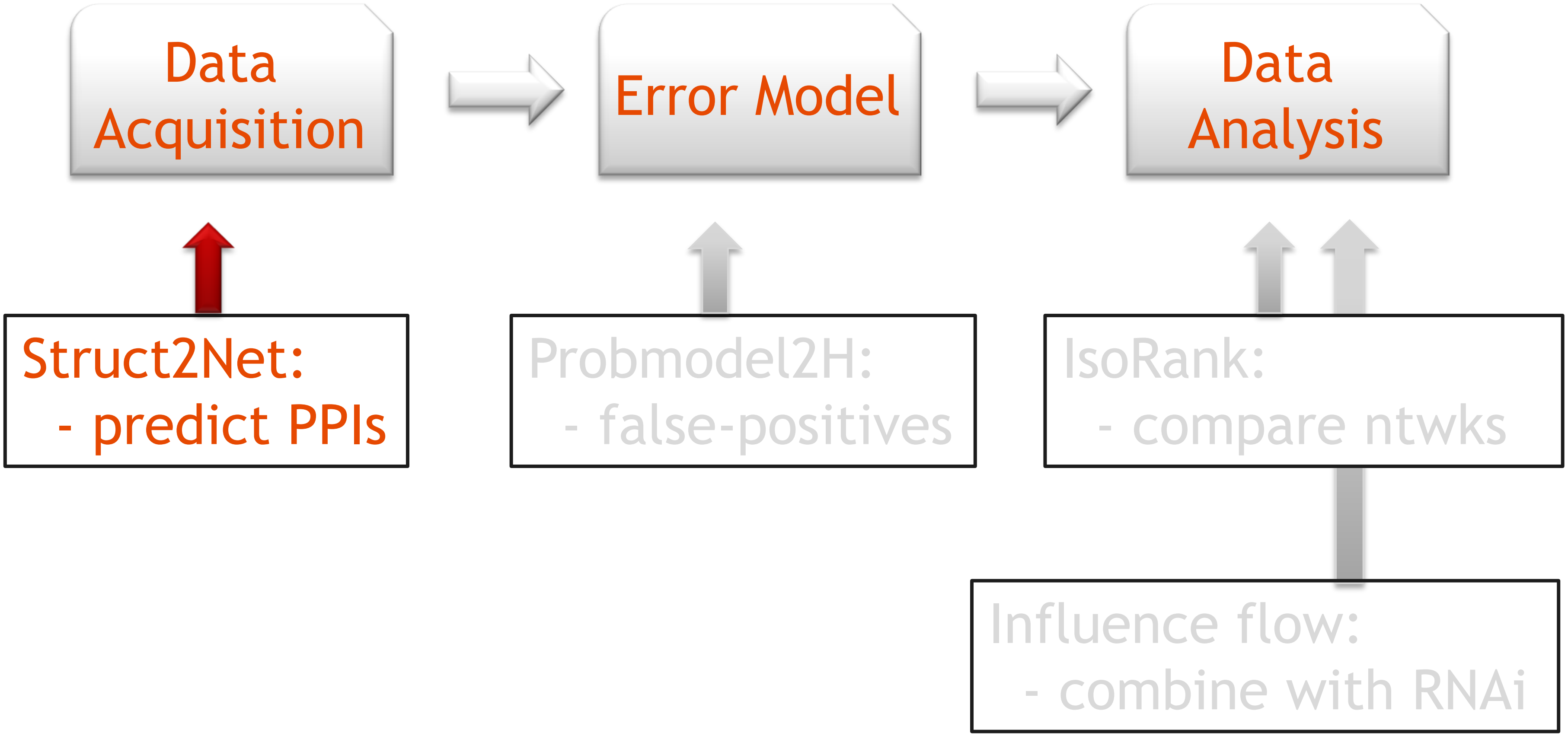


Results: using full MAPK cascade



Outline

- Introduction to Protein Interactions
- Algorithms for PPI Networks:



Data Acquisition

Error Model

Data Analysis

Struct2Net:
- predict PPIs

Probmodel2H:
- false-positives

IsoRank:
- compare ntwks

Influence flow:
- combine with RNAi

Struct2Net

Goal: computationally predict if two proteins physically interact

Why?

- Prune the list of interactions to test
- Help identify experimental errors

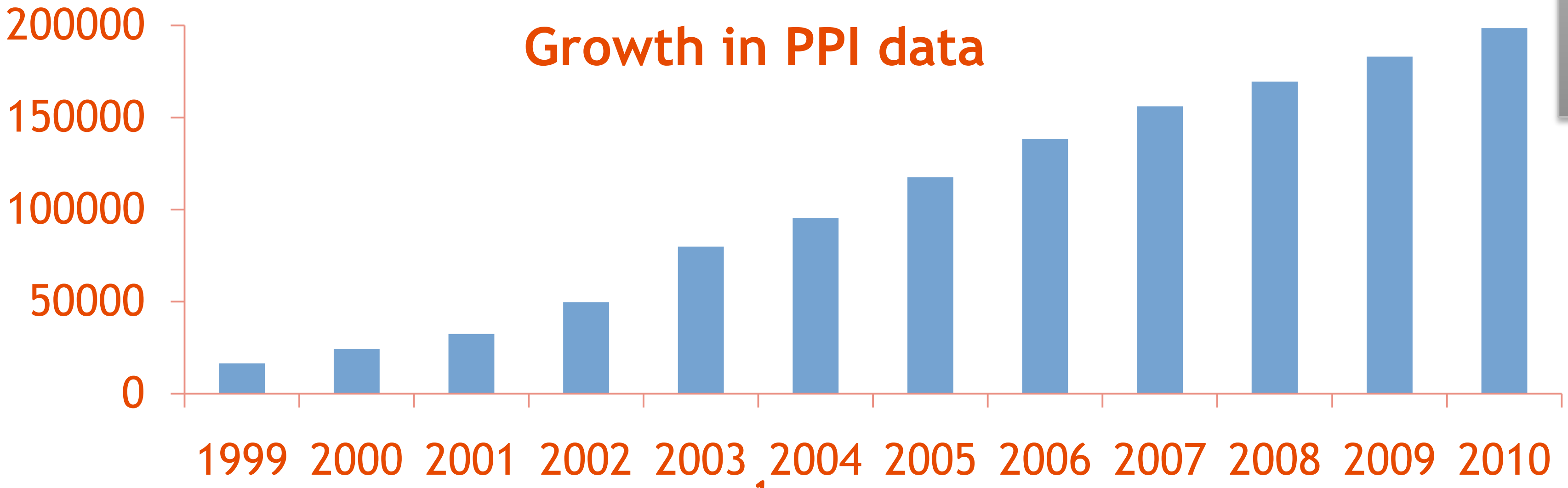
How?

- Use ideas from structural biology
- Machine Learning approach: pose as a classification task

Acknowledgments

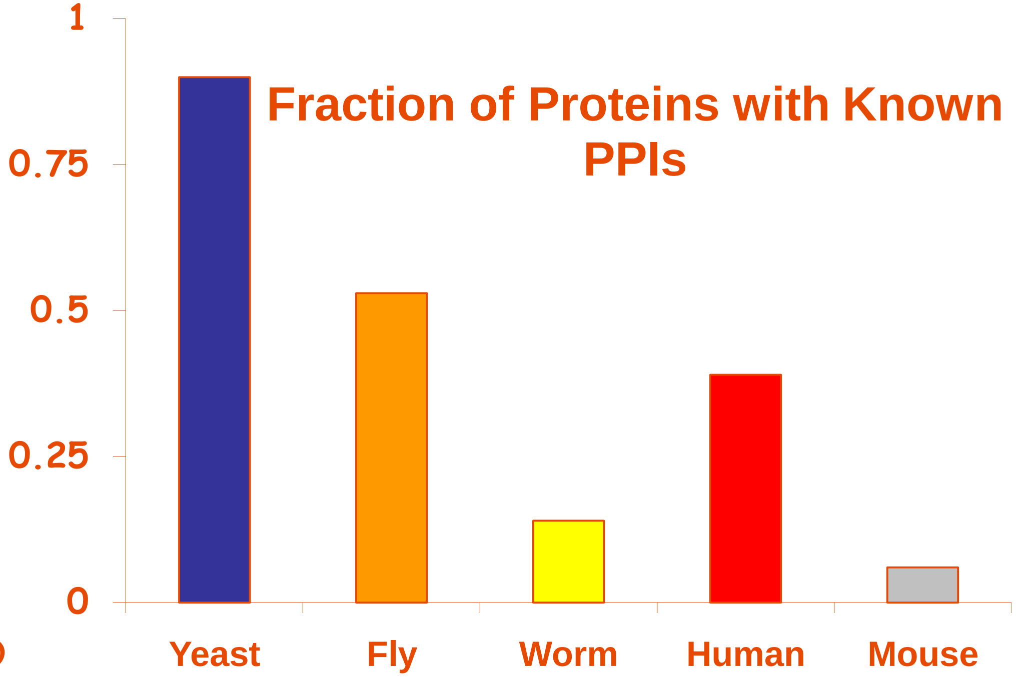
- Collaborators:
  - Struct2Net: Jinbo Xu & Bonnie Berger
  - Struct2Net-DB: Daniel Park, Jinbo Xu, Raghu Hosur & Bonnie Berger
- Previously presented/published:
  - PSB 2006
  - Nucleic Acids Research (Web Server Issue), 2010

Why: the data is not nearly enough…



Main problems:

- O(n^2): Too many possible interactions
- High-throughput methods are error-prone



PPI statistics based on data from BIOGRID

Problem Formulation

Given

1. two protein sequences
2. a database of protein-complex structures
3. [Optional] measures of functional relationships between the two proteins

Find

probability of interaction between the two proteins

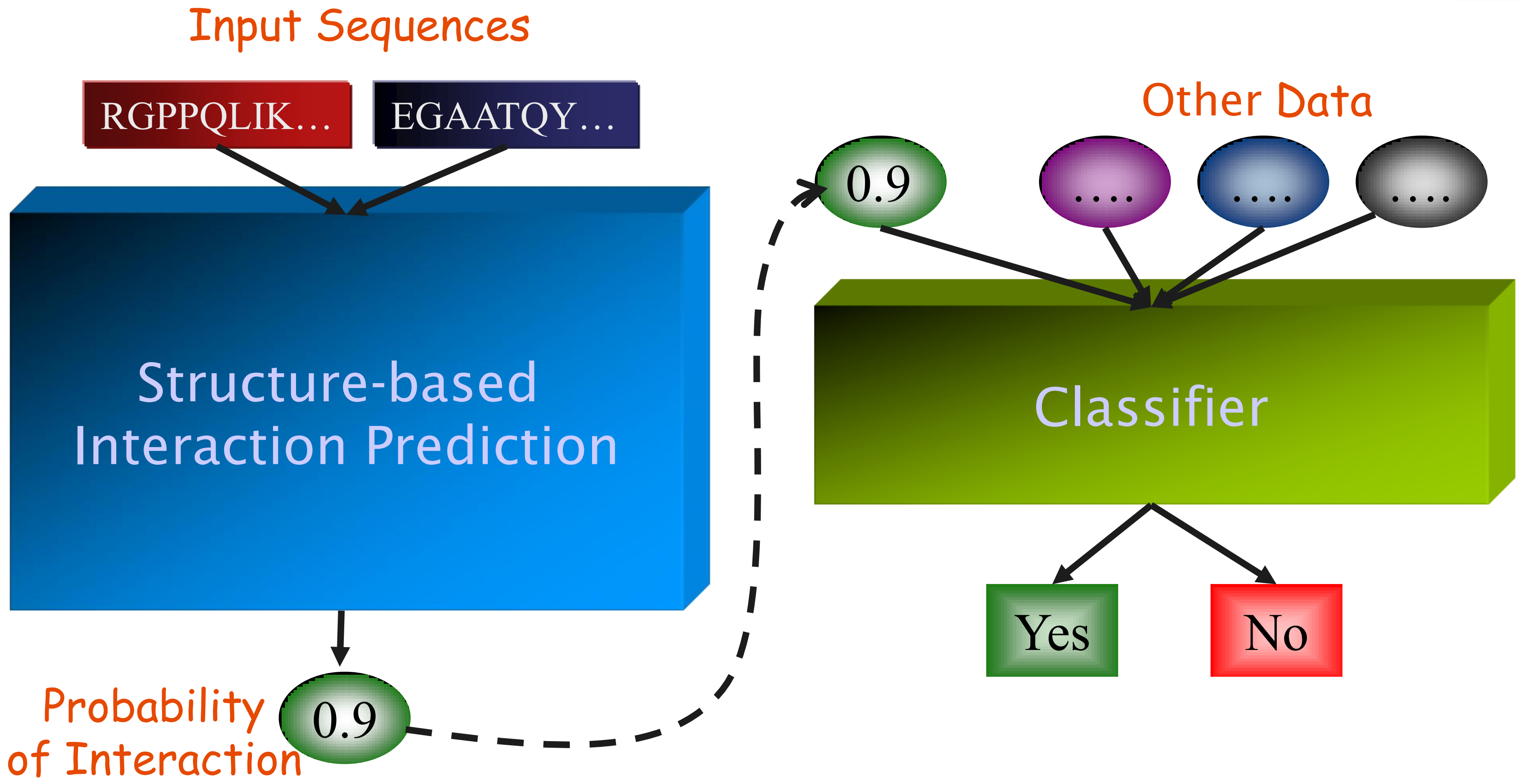
Evaluation

1. Using known PPI data, construct datasets of high-confidence positive and negative examples
2. Estimate predictive power on this dataset

Previous Approaches vs. Us

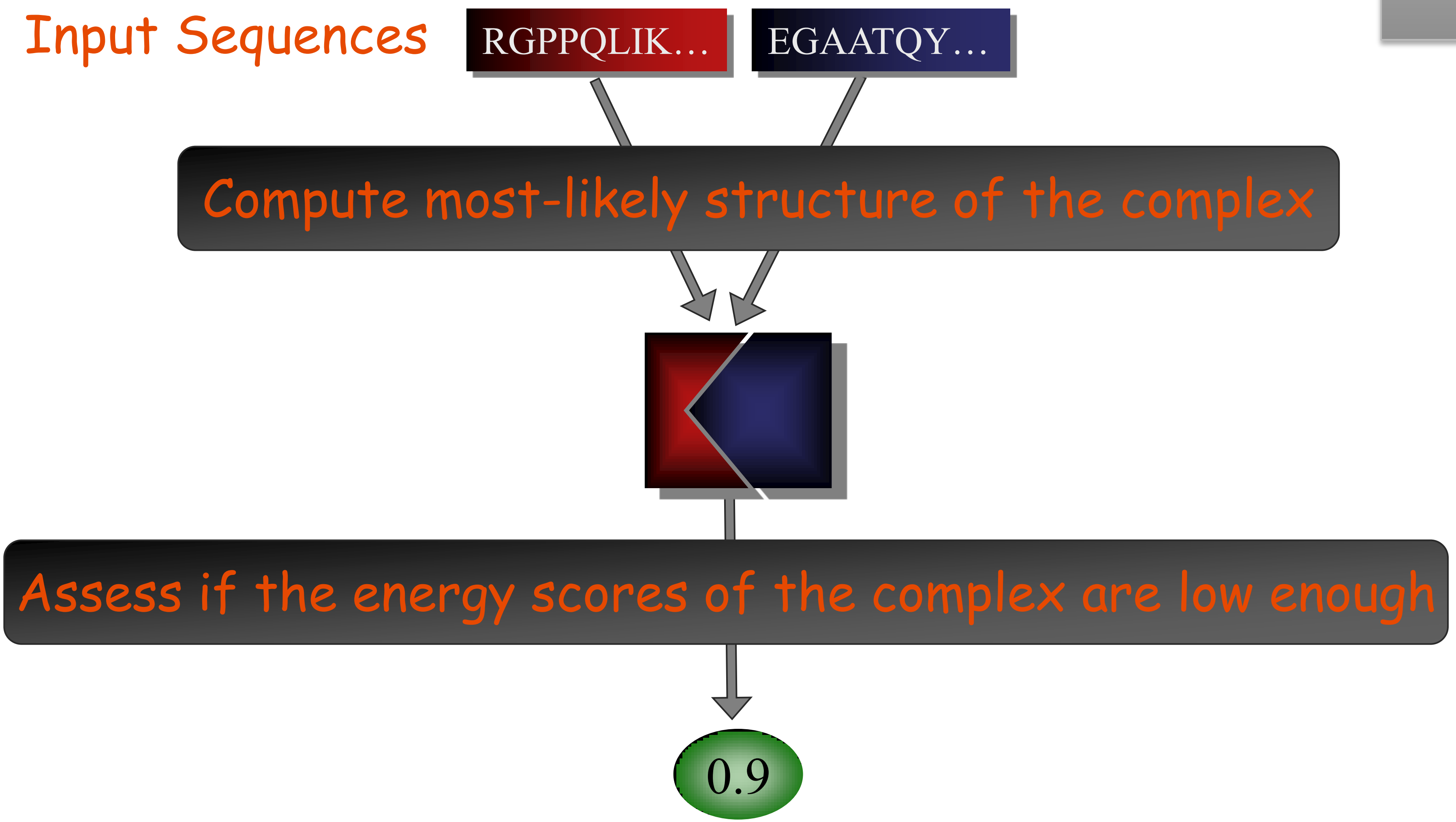
- Guilt by association: proteins that interact often have similar functional characteristics
  - Pose as a classification problem.
  - Missing data issues
- Biological models: correlated mutations, sequence domains
- We use a structure-based approach:
  - Can figure out why/how an interaction happens
  - Works even when functional data is unavailable

Outline of Our Approach



Predicting Interaction Using Structure

Struct2Net



Input Sequences

RGPPQLIK…

EGAATQY…

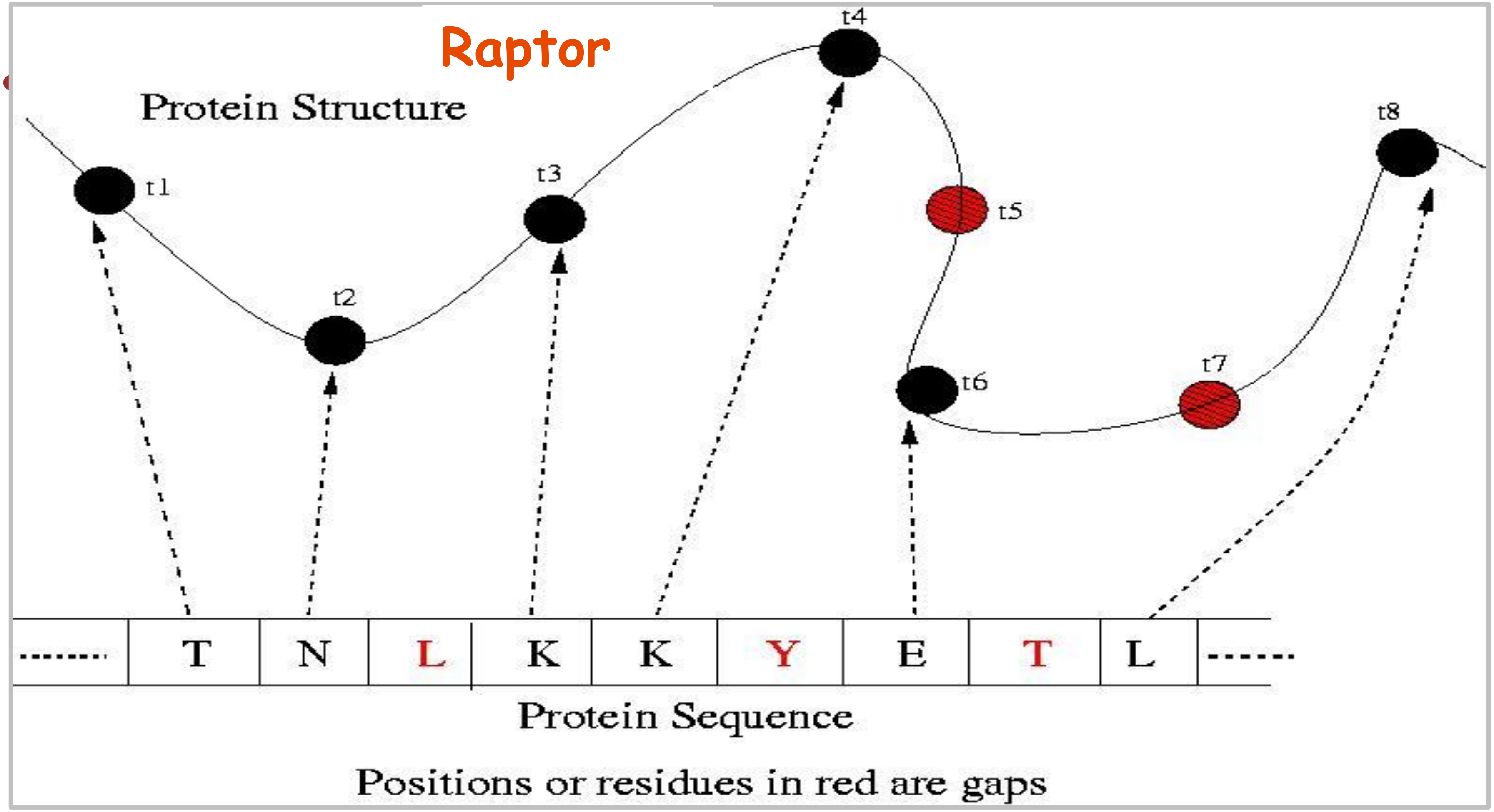
Compute most-likely structure of the complex

Assess if the energy scores of the complex are low enough

0.9

Joint Homology Modeling

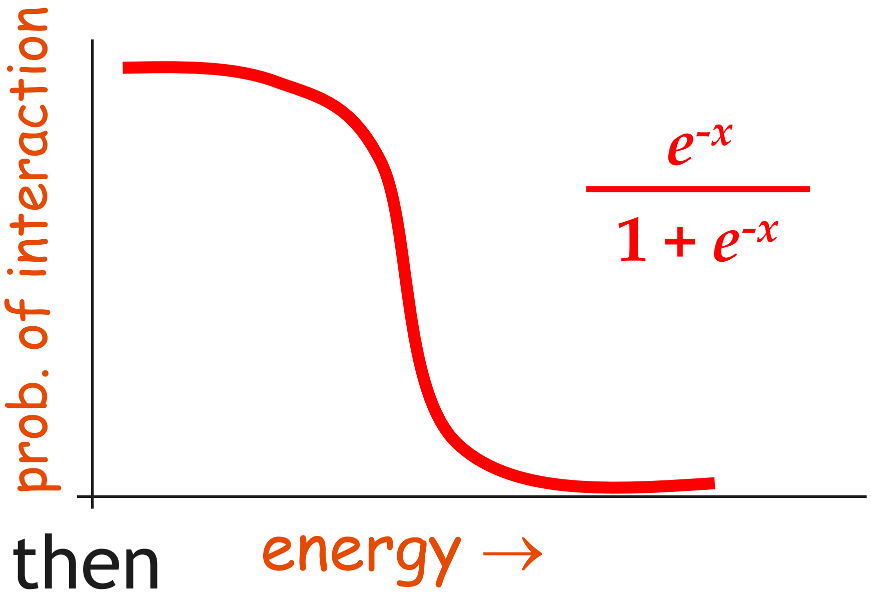
- Goal: Find optimal alignment of sequence to template structure



Raptor

Energy Scores → Interaction Probability

- Want to summarize multiple energy scores into one probability score
- Logistic Regression



$S_1 \ldots S_K$ are energy scores, then

$$\boxed{P(\text{ interact } \mid S_1 \ldots S_K) = \text{logit}(a_1 S_1 + \ldots + a_K S_K)}$$

$$\text{where, logit}(x) = \frac{1}{1 + e^{-x}}$$

Model Selection: which features to use

- We tried various combinations of energy scores, including normalized-energy scores to the set of parameters

$$S_{\text{normalized}} = \frac{S}{\text{mean sequence length}}$$

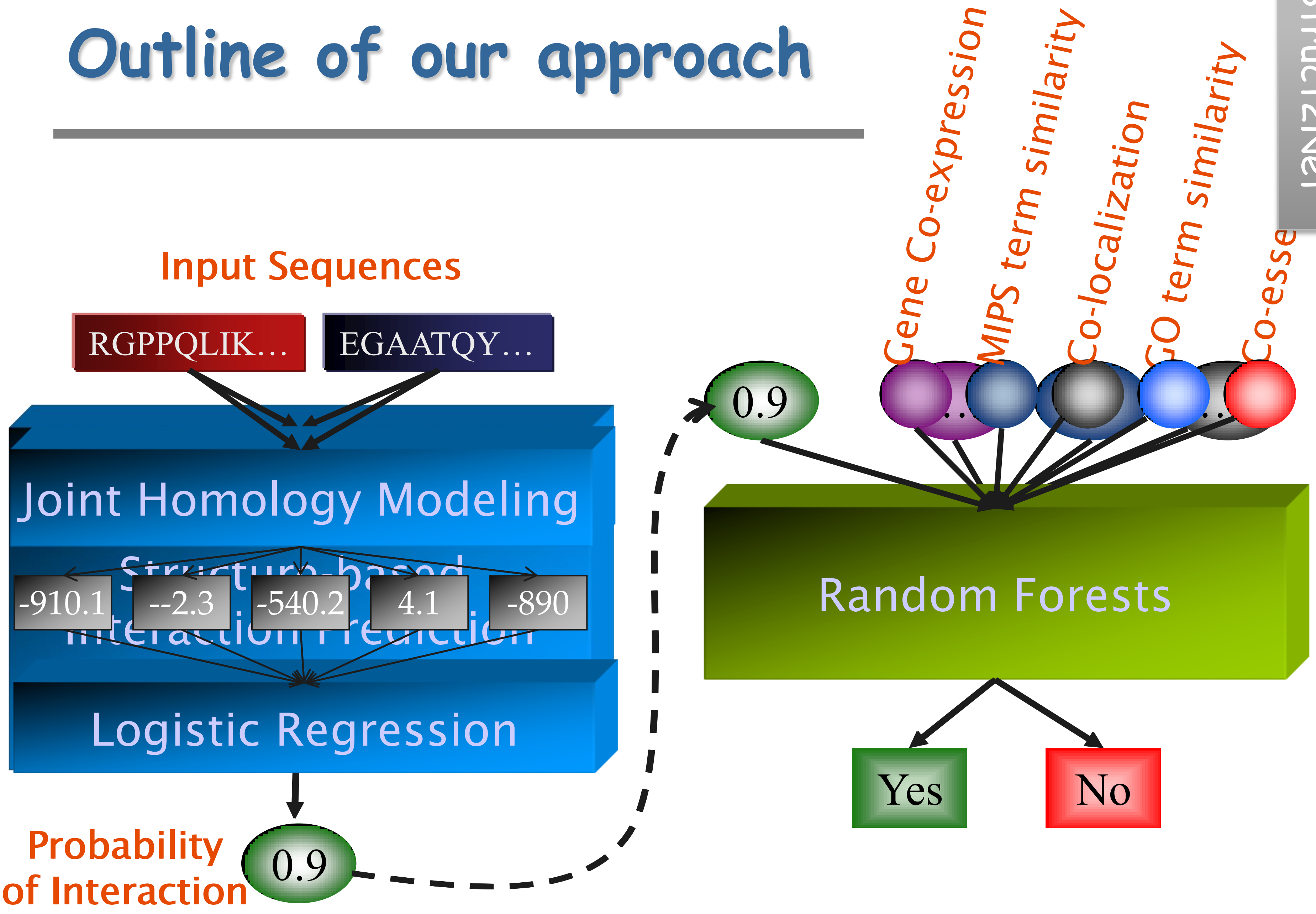
- Model selection to identify the best predictors
  - AIC based feature selection
  - L1-norm regularized logistic regression

$$\min_\theta \sum -\log(p(y|\mathbf{x};\theta)) \quad \rightarrow \quad \min_\theta \sum -\log(p(y|\mathbf{x};\theta)) + \beta|\theta|_1$$

- Normalized energy and alignment scores win over raw scores

Struct2Net

Outline of our approach



Input Sequences

RGPPQLIK…

EGAATQY…

Joint Homology Modeling

Structure-based Interaction Prediction

-910.1

--2.3

-540.2

4.1

-890

Logistic Regression

Probability of Interaction

0.9

0.9

Gene Co-expression

MIPS term similarity

Co-localization

GO term similarity

Co-esse

…

…

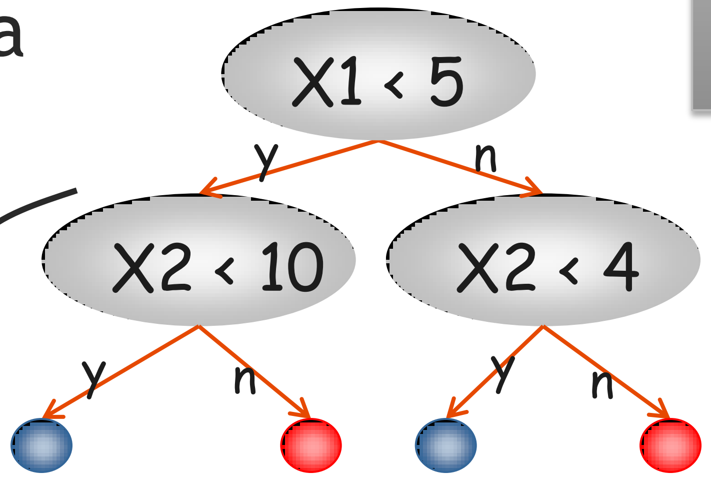
Random Forests

Yes

No

Random Forests

- Extend the decision tree idea

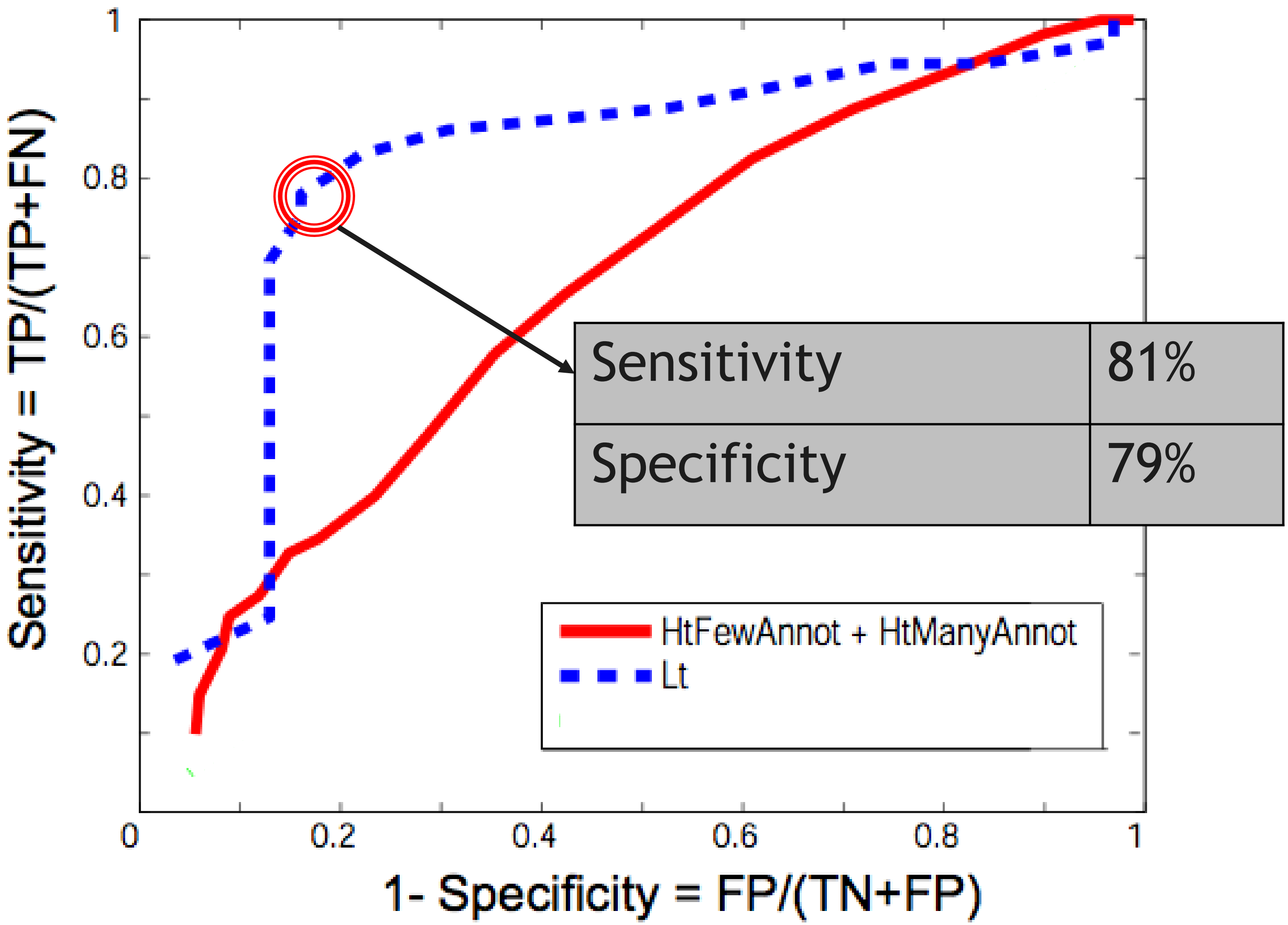


What if the value along X2 is not known ?

- Make many trees:
  - Each trained on only a subset of features
  - To classify a new point, take majority vote

	X1	X2	X3	X4	X5
T1	🟩	⬜	🟩	🟩	🟩
T2	🟩	🟩	⬜	🟩	⬜
T3	⬜	🟩	🟩	🟩	🟩

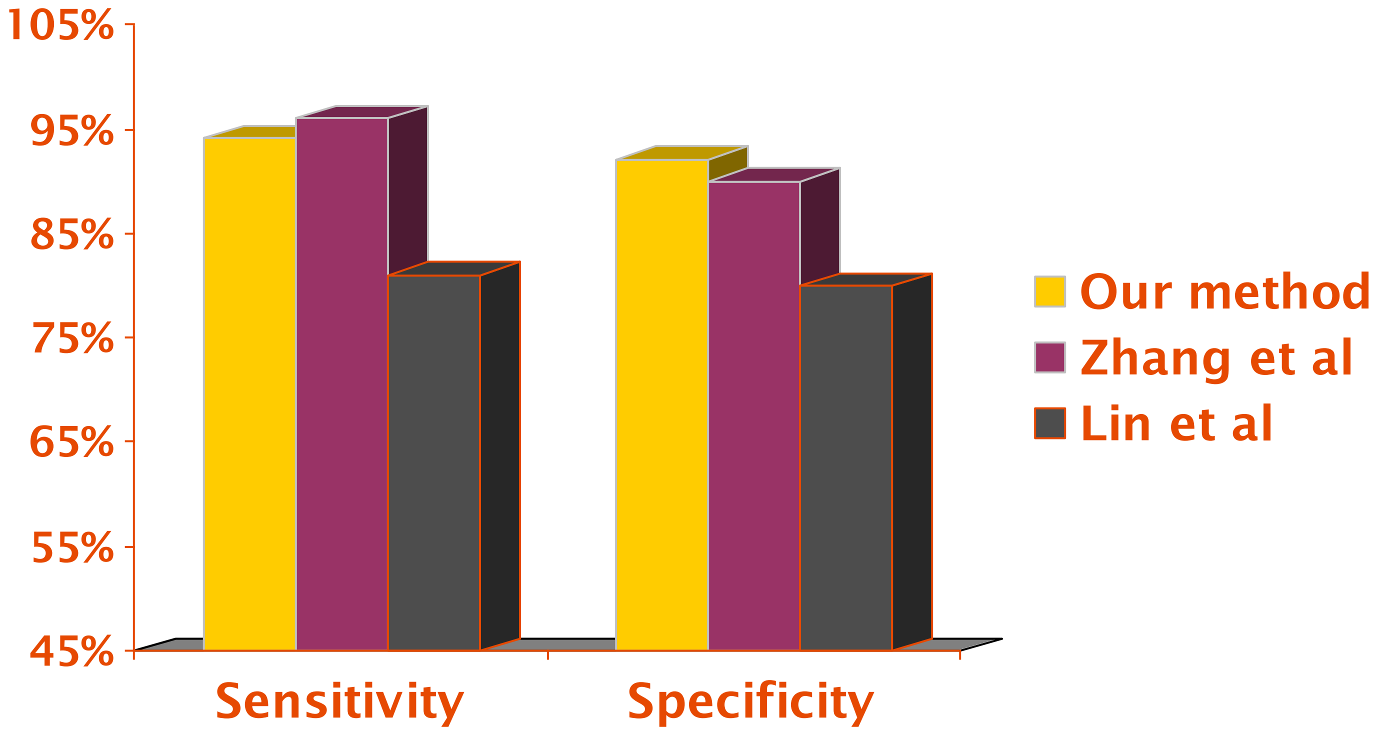
Using only Structure-based Method



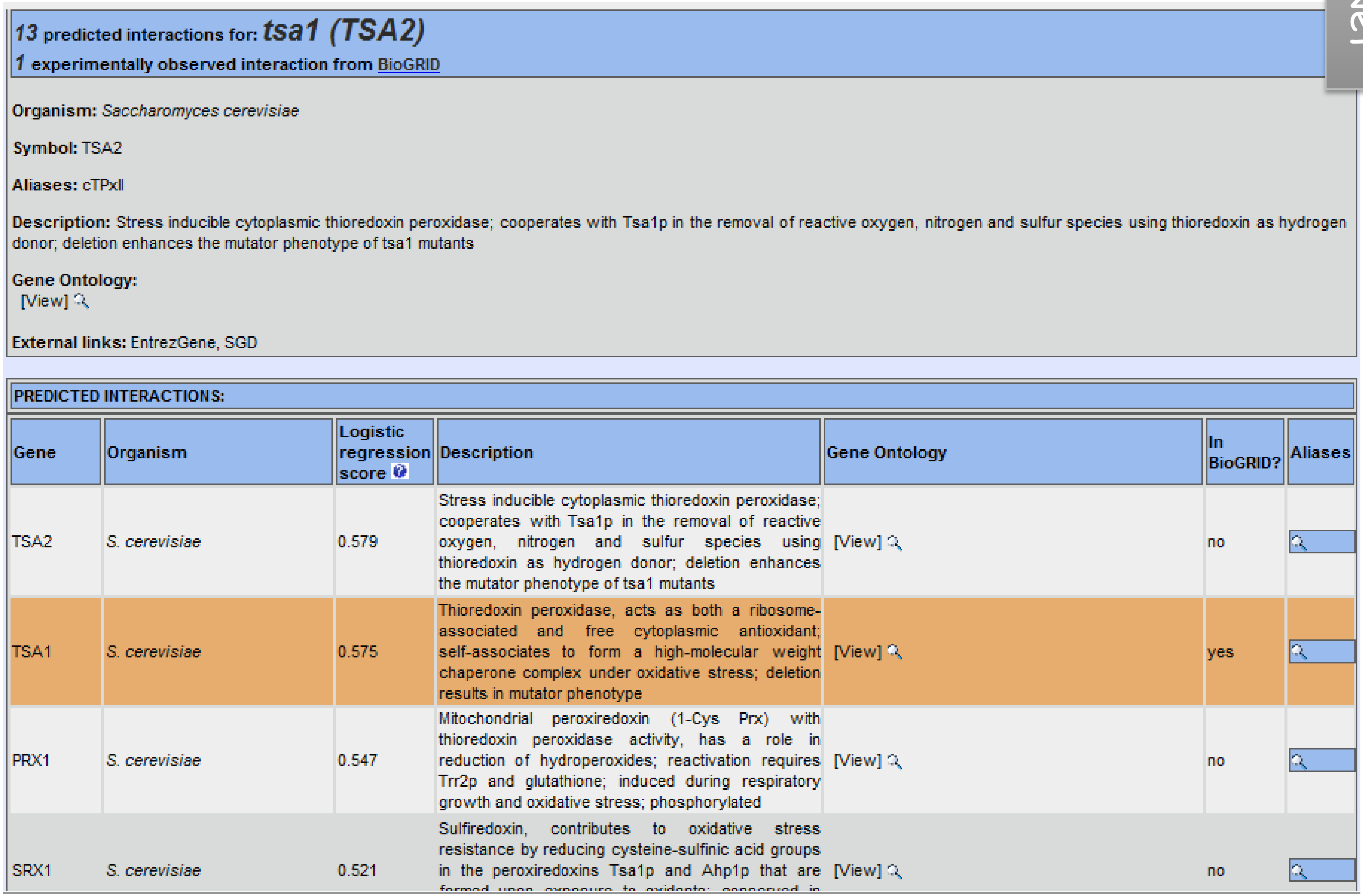
Sensitivity	81%
Specificity	79%

Structure + Other Information

Comparison with *Lin et al, BMC Bioinfo., 2004*



Struct2Net DB



13 predicted interactions for: *tsa1 (TSA2)*
1 experimentally observed interaction from BioGRID

Organism: *Saccharomyces cerevisiae*

Symbol: TSA2

Aliases: cTPxII

Description: Stress inducible cytoplasmic thioredoxin peroxidase; cooperates with Tsa1p in the removal of reactive oxygen, nitrogen and sulfur species using thioredoxin as hydrogen donor; deletion enhances the mutator phenotype of tsa1 mutants

Gene Ontology:
[View] 🔍

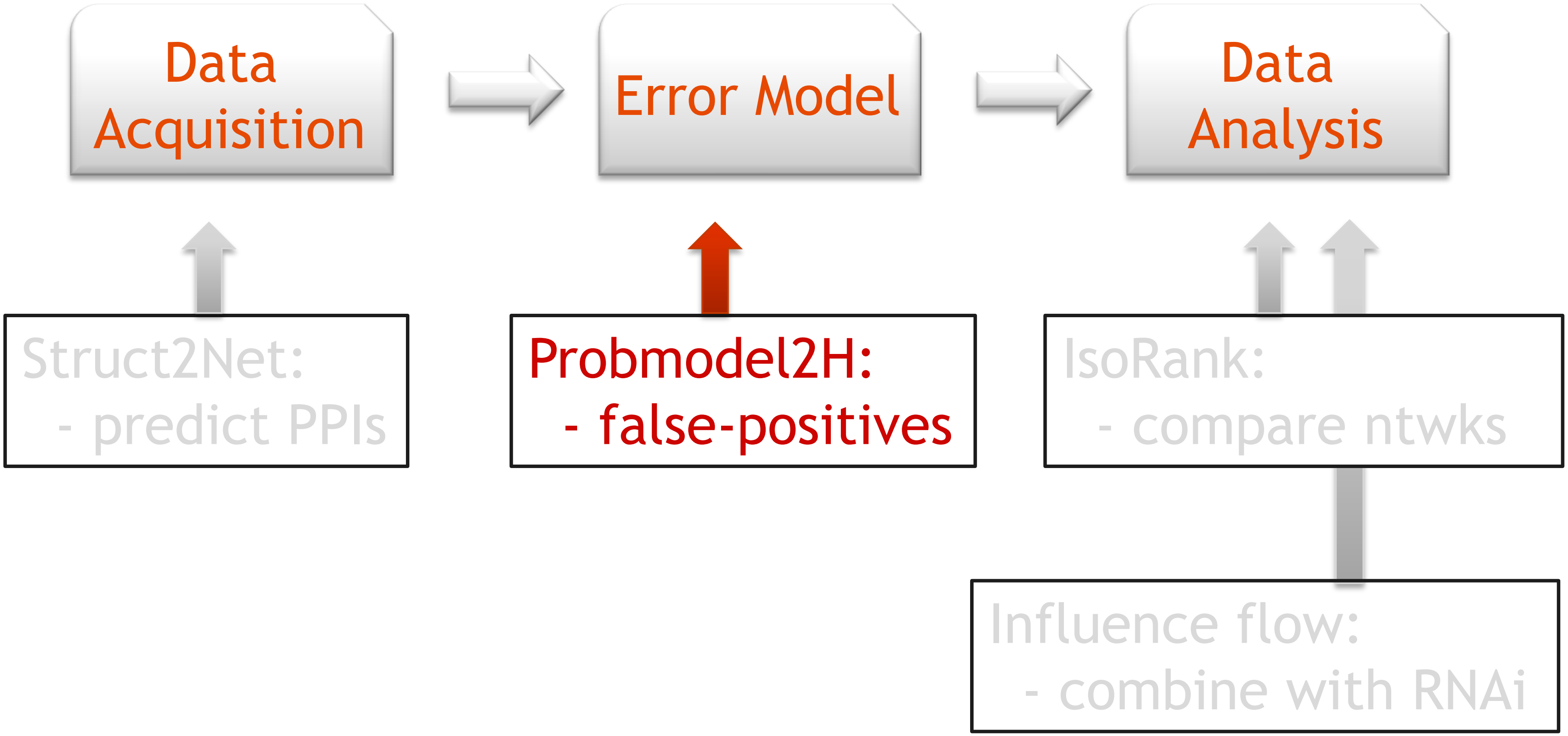
External links: EntrezGene, SGD

PREDICTED INTERACTIONS:

Gene	Organism	Logistic regression score ?	Description	Gene Ontology	In BioGRID?	Aliases
TSA2	*S. cerevisiae*	0.579	Stress inducible cytoplasmic thioredoxin peroxidase; cooperates with Tsa1p in the removal of reactive oxygen, nitrogen and sulfur species using thioredoxin as hydrogen donor; deletion enhances the mutator phenotype of tsa1 mutants	[View] 🔍	no	🔍
TSA1	*S. cerevisiae*	0.575	Thioredoxin peroxidase, acts as both a ribosome-associated and free cytoplasmic antioxidant; self-associates to form a high-molecular weight chaperone complex under oxidative stress; deletion results in mutator phenotype	[View] 🔍	yes	🔍
PRX1	*S. cerevisiae*	0.547	Mitochondrial peroxiredoxin (1-Cys Prx) with thioredoxin peroxidase activity, has a role in reduction of hydroperoxides; reactivation requires Trr2p and glutathione; induced during respiratory growth and oxidative stress; phosphorylated	[View] 🔍	no	🔍
SRX1	*S. cerevisiae*	0.521	Sulfiredoxin, contributes to oxidative stress resistance by reducing cysteine-sulfinic acid groups in the peroxiredoxins Tsa1p and Ahp1p that are formed upon exposure to oxidants; conserved in	[View] 🔍	no	🔍

Outline

- Introduction to Protein Interactions
- Algorithms for PPI Networks:



Data Acquisition → Error Model → Data Analysis

Struct2Net:
- predict PPIs

Probmodel2H:
- false-positives

IsoRank:
- compare ntwks

Influence flow:
- combine with RNAi

Goal: identify false-positives in Yeast 2Hybrid data

Why?

- Systematic false positives can occur
  - *“at times, the functional co-relevance of two proteins scored as interacting in the two-hybrid system is unlikely.”* (Serebriiskii et al, Biotechniques, 2000)
  - *“Y2H screens suffer …. from false positives, i.e. interactions that appear to take place only in the context of the Y2H assay”* (Stellberger et al, Protein Science, 2010)

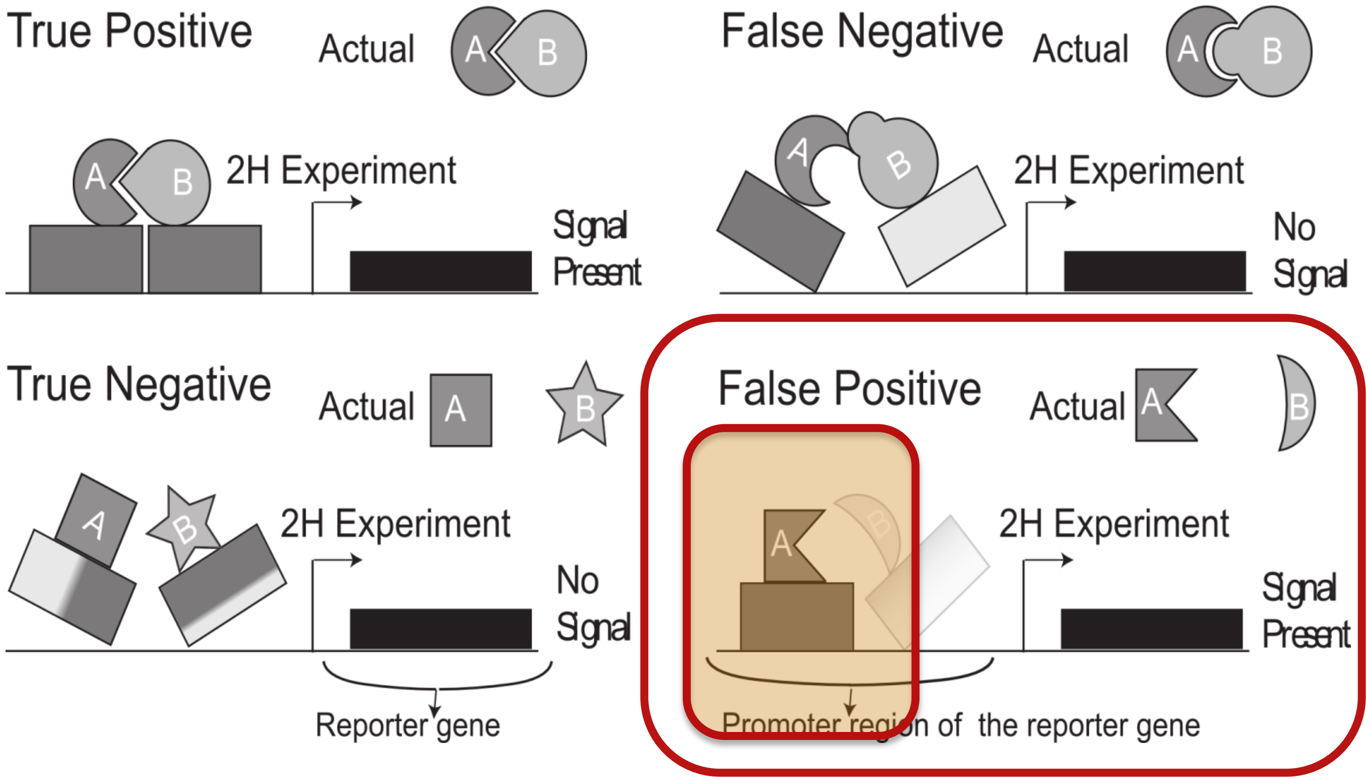
How?

- Bayesian model to identify “promiscuous” proteins

Acknowledgments

- Collaborators:
  - David Sontag & Bonnie Berger
- Previously presented/published:
  - PSB 2007

Errors in Y2H experiments



Problem Formulation

Probmodel 2H

Given

1. Datasets D_1, D_2, … of Y2H data for a single species, each from a single experimental setup. Each D_i is a list of protein-pairs.
2. [Optional] For some dataset D_i, a score indicating confidence in each data-point in D_i
3. [Optional] Other datasets (e.g. from Literature) indicating interaction between proteins in the species

Find

1. for each protein-pair, probability of true interaction
2. for each protein, an estimate of its Y2H promiscuity

Evaluation

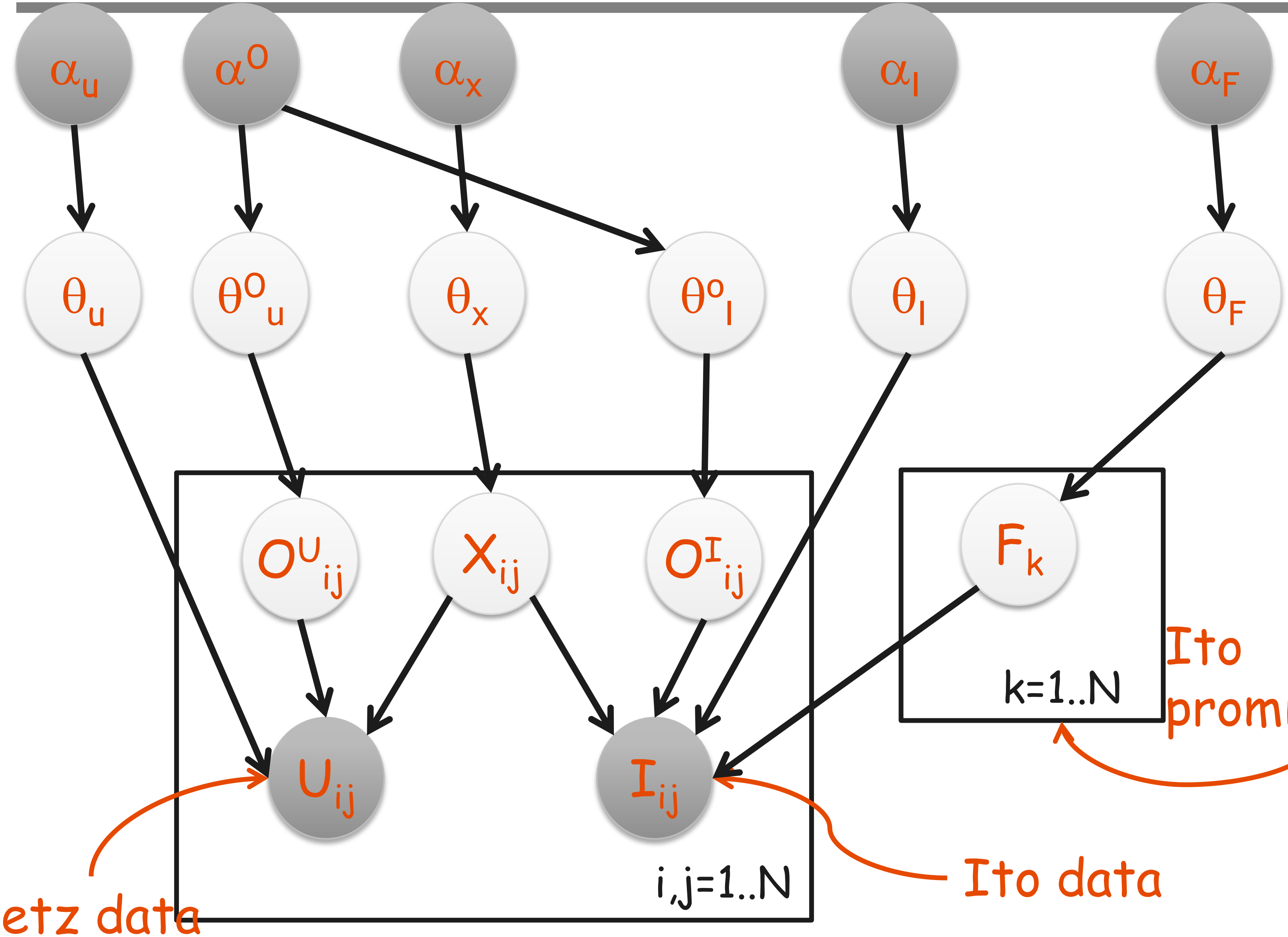
1. Using known Y2H and CoIP PPI data, construct datasets of high-confidence positive and negative examples of Y2H PPIs
2. Estimate predictive power on this dataset

Previous Work vs. Us

- Some previous approaches:
  - Require overlap between Y2H & Co-IP data
  - Use repetition data from each experiment
  - Product of node-degrees (Bader et al.)
- Us:
  - Set up a Bayesian framework to identify promiscuous proteins
  - Can learn across multiple datasets

Initial approach: Generative Model

Probmodel 2H

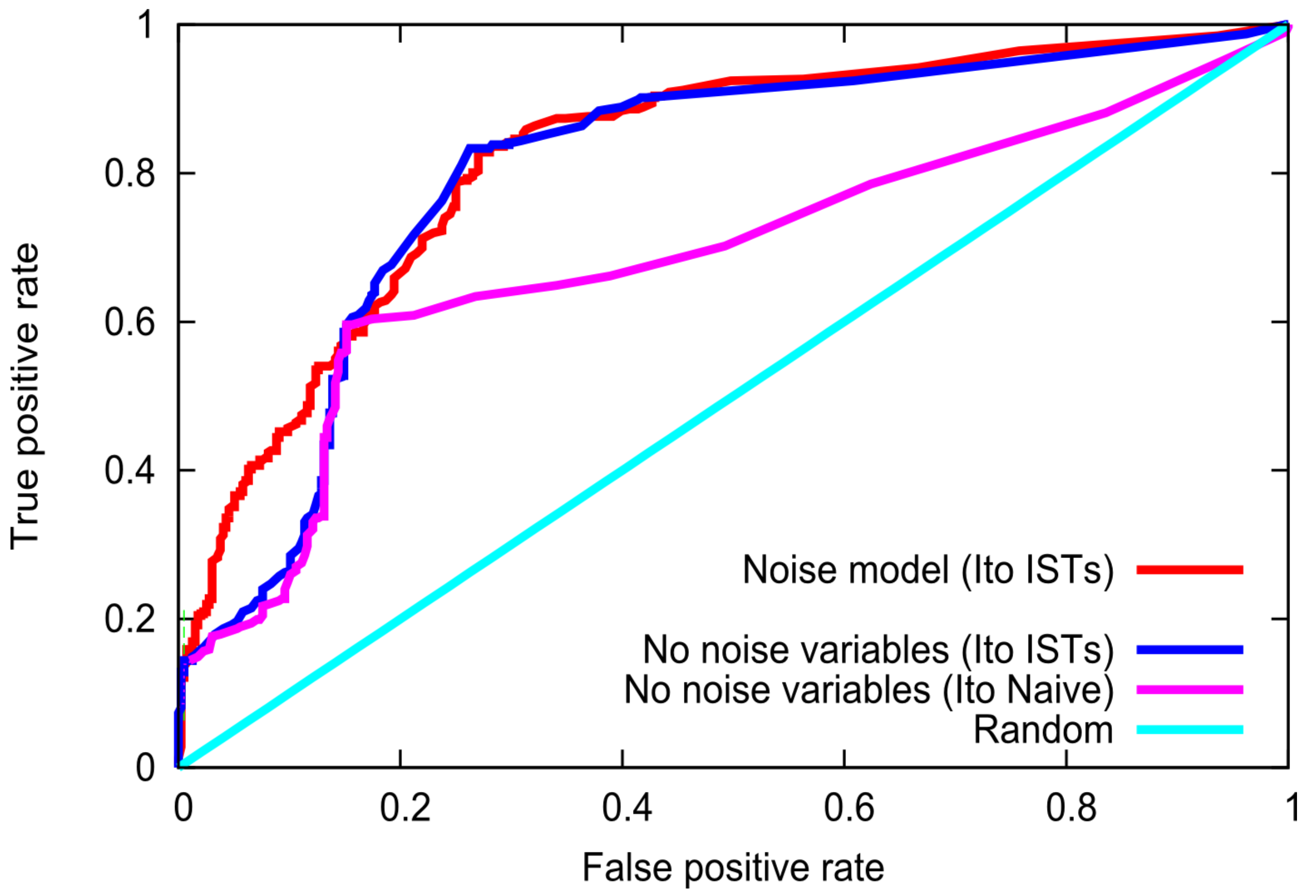


Uetz data

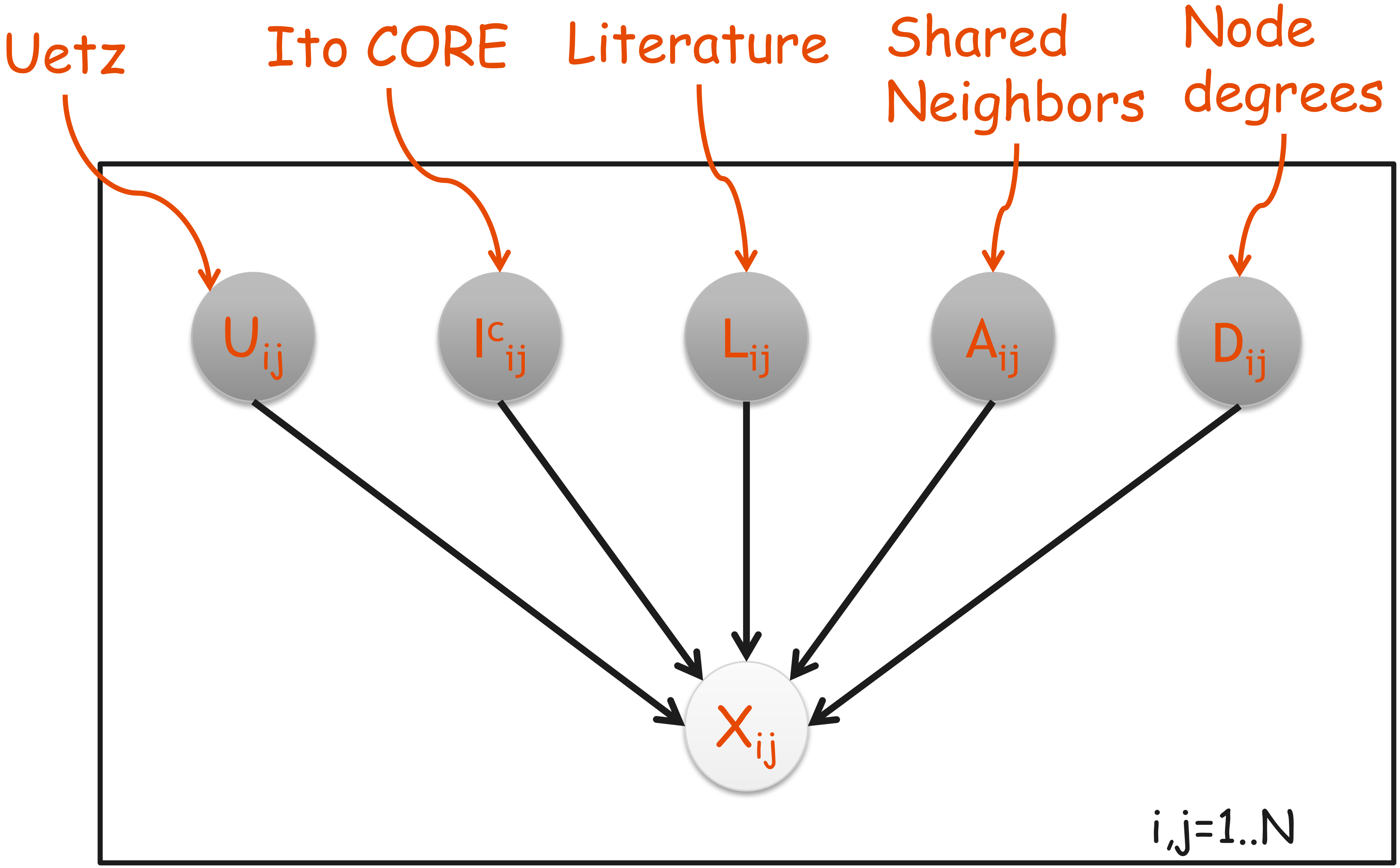
Ito data

Ito promiscuous

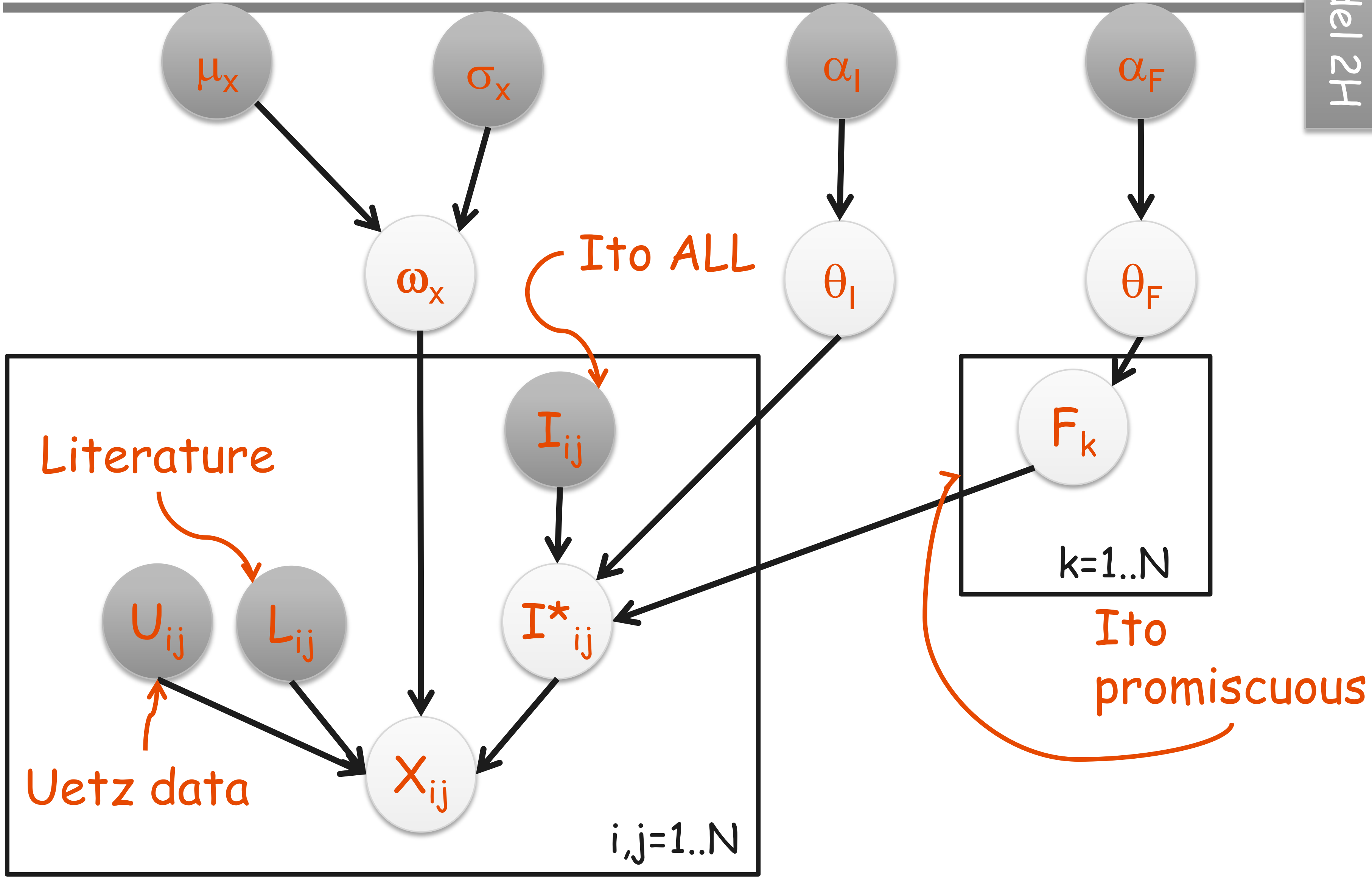
Results: Generative Model



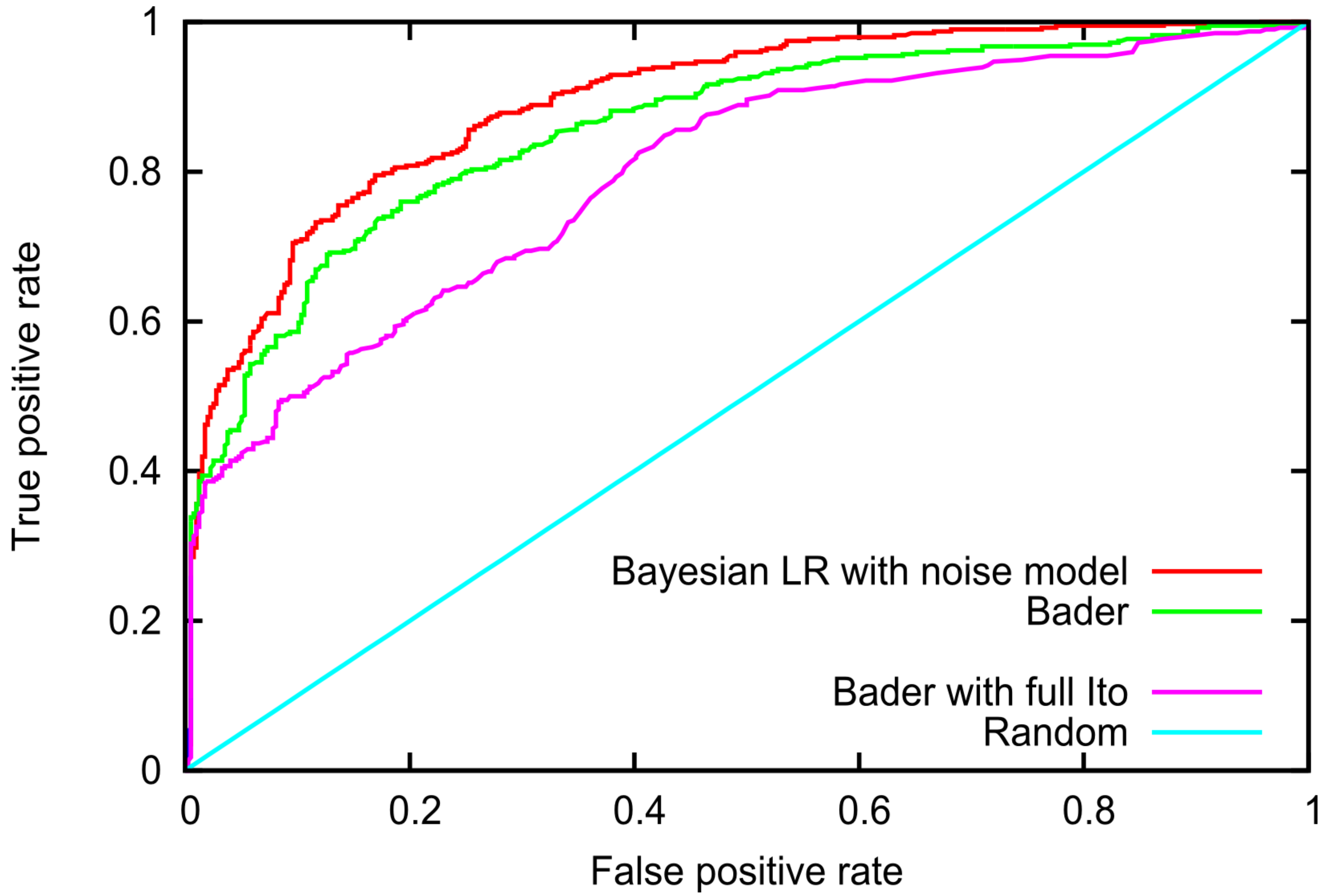
Logistic Regression Approach: Bader et al.



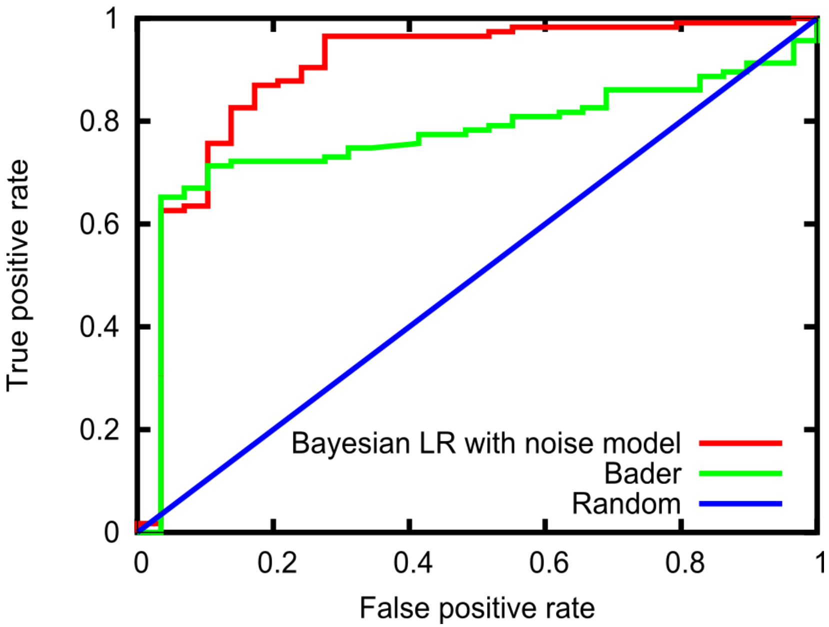
Our Logistic Regression Model



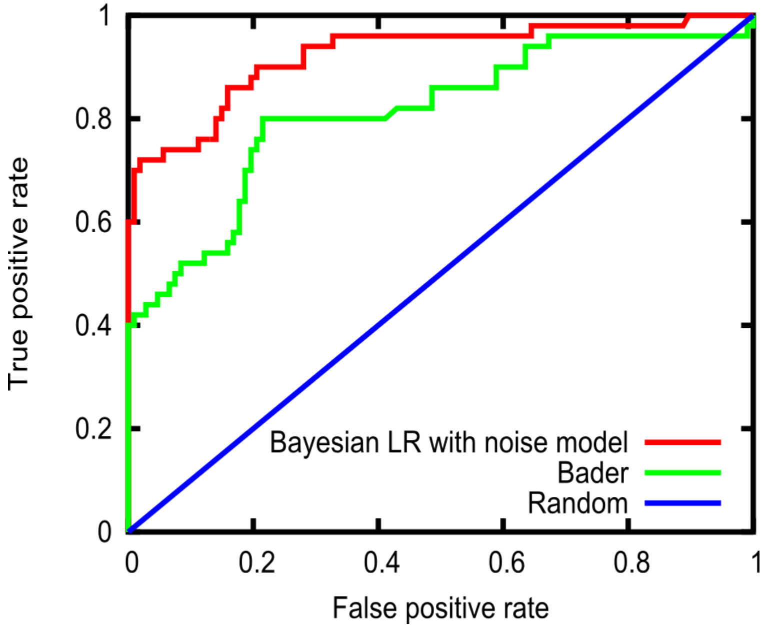
Results: Logistic Regression Models



The Bayesian Model Really Helps in Certain Cases



Medium degree with positive hit in Uetz or Literature



High degree

We Get More Fine-grained Promiscuity Estimates

Protein	Degree	P(promiscuous)
YJR091C	285	0.389
YMR047C	125	0.481
YLR295C	124	0.513
YNL189W	122	0.5
YPR086W	99	0.492
YER022W	98	0.253
YER081W	95	0.486
YHR114W	91	0.491
YLR447C	88	0.498
YLR453C	79	0.498
YLR288C	78	0.498

Protein	Degree	P(promiscuous)
YGL127C	68	0.125
YDR034C	63	0.495
YLR423C	60	0.373
YML064C	54	0.516
YGL070C	44	0.435
YKL002W	40	0.484
YDR318W	34	0.297
YGR218W	34	0.182
YDL153C	32	0.274
YLR373C	31	0.457
YPL070W	30	0.492

Thanks!

- Bonnie Berger
- Dave Gifford & Srini Devadas
- Patrice Macaluso
- Berger Group: Allen, Andrew, Beckett, Charlie, Danny, George, Irene, Jinbo, Leonid, Luke, Michael, Mike, Nathan, Patrick, Shannon…
- Perrimon Lab @ HMS: Adam Friedman, Chris Bakal, Norbert Perrimon