CS4220: Knowledge Discovery Methods for Bioinformatics

Unit 5: Biological Network

Limsoon Wong

NUS National University of Singapore

Lecture Outline

- Overview of biological networks
- Use of biological networks in enhancing bioinformatics analysis
- Consistency, comprehensiveness, and compatibility of biological pathway databases
- Integrating pathway databases
- Reliability of PPIN
- Identifying noise edges in PPIN
- Identifying missing edges in PPIN
- An advanced example on assessment of PPIN

Overview of Biological Networks

Why Biological Networks?

- Complete genomes are now available
- Knowing the **genes** is not enough to understand how biology **functions**

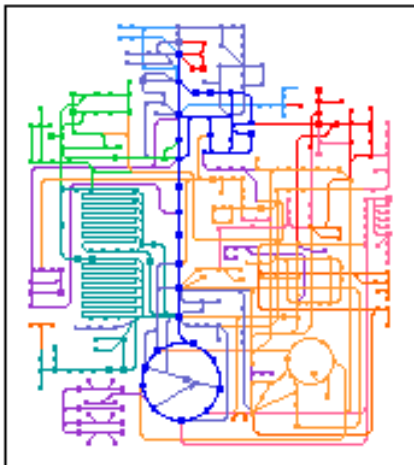
- **Proteins,** not genes, are responsible for many cellular activities

- Proteins function by **interacting** w/ other proteins and biomolecules

GENOME

PROTEOME

"INTERACTOME"



Slide credit: See-Kiong Ng

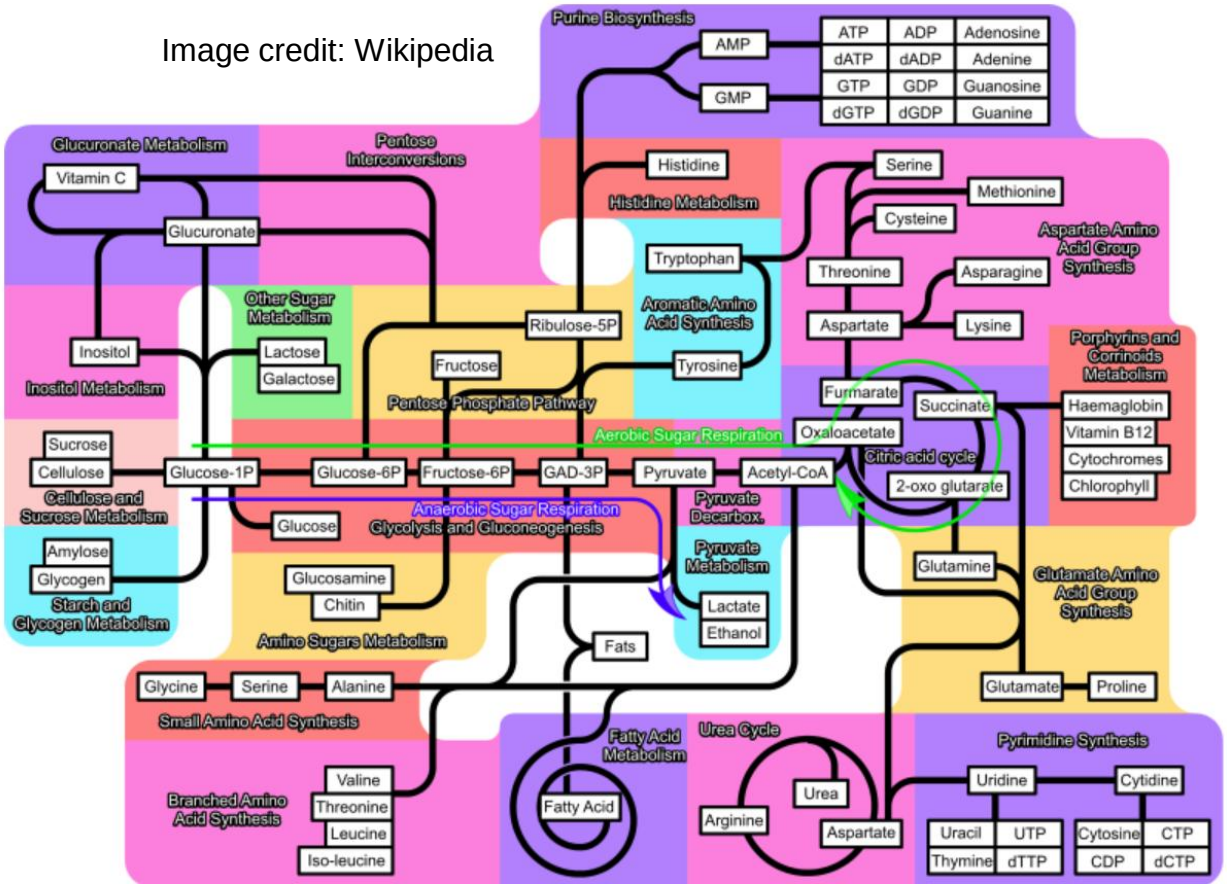
Types of Biological Networks

- **Natural biological pathways**
  - Metabolic pathway
  - Gene regulation network
  - Cell signaling network

- **Protein-protein interaction networks**

Metabolic Pathway

Image credit: Wikipedia



- **A series of biochem reactions in a cell**
  - Catalyzed by enzymes
  - Step-by-step modification of an initial molecule to form another product that can
    - **be used /store in the cell**
    - **initiate another metabolic pathway**

Gene Regulation Network

- **Gene regulation is the process that turns info from genes into gene products**
- **Gives a cell control over its structure & function**
  - Cell differentiation
  - Morphogenesis
  - Adaptability, …

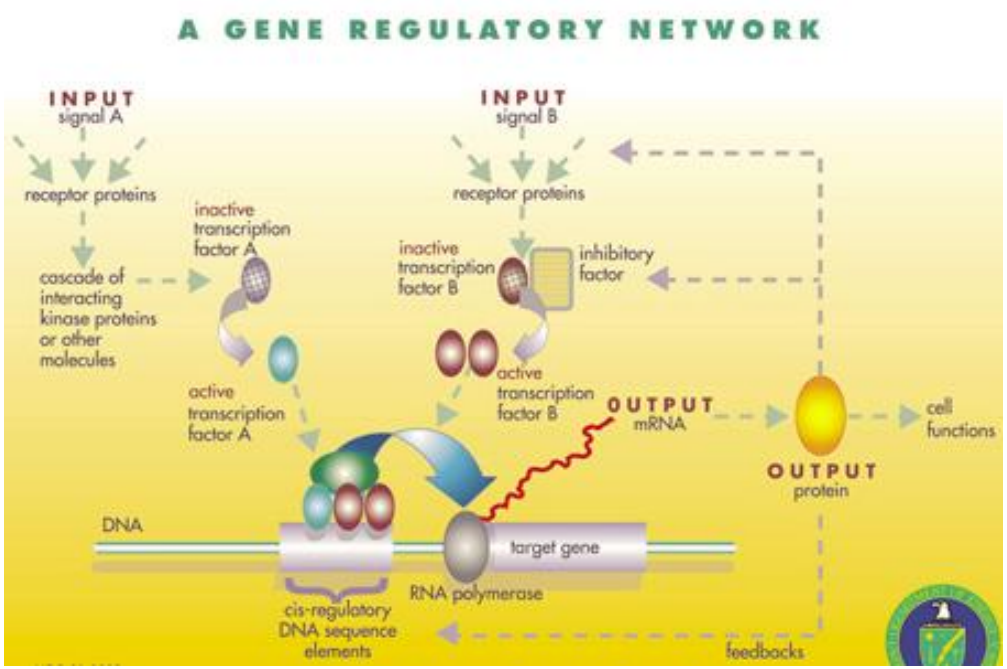


Image credit: Genome to Life

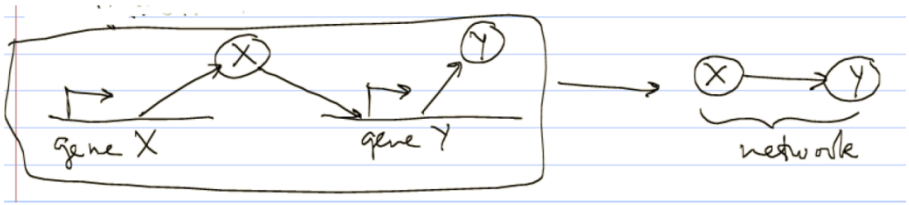


Image credit: Natasa Przulj

Cell Signaling Network

- **It is the entire set of changes induced by receptor activation**
  - Governs basic cellular activities and coordinates cell actions
- **Cells communicate with each other**
  - Direct contact (juxtacrine signaling)
  - Short distances (paracrine signaling)
  - Large distances (endocrine signaling)
- **Errors result in cancer, diabetes, ...**

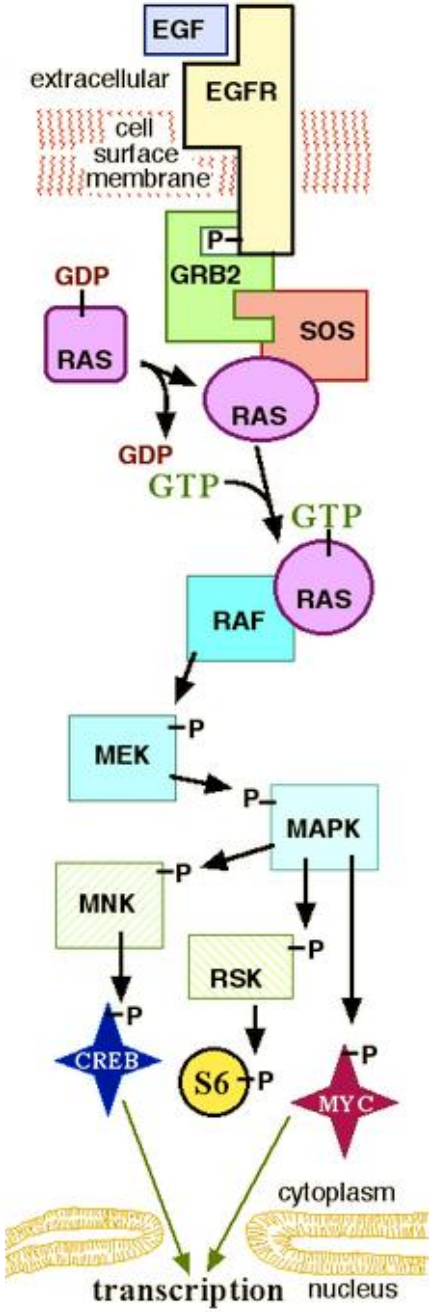
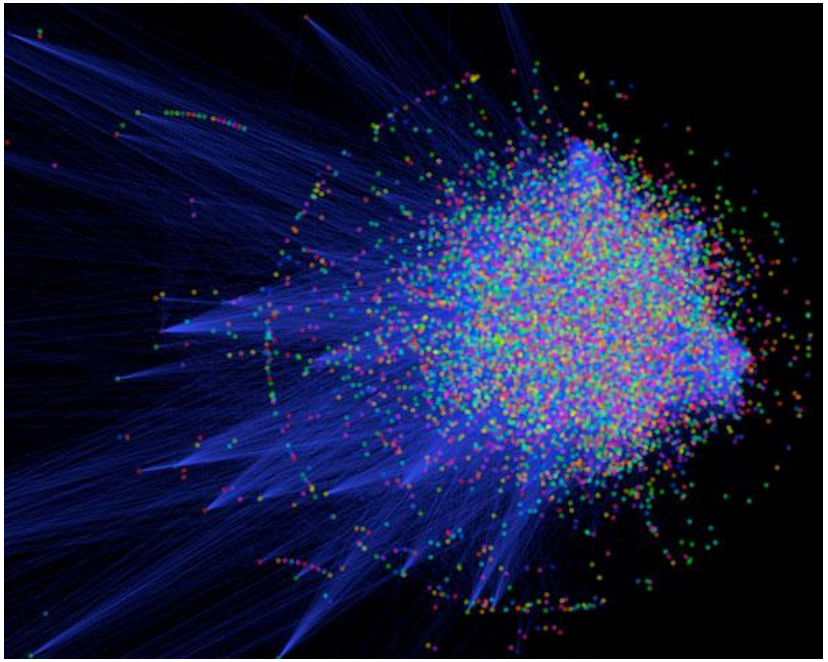


Image credit: Wikipedia

Protein Interaction Network (PPIN)

- **PPI usual refers to physical binding between proteins**
  - Stable interaction
    - **Protein complex**
    - **~70% of PPIs**
  - Transient interaction, modifying a protein for further actions
    - **Phosphorylation**
    - **Transportation**
    - **~30% of PPIs**



Visualization of the human interactome.
Image credit: Wikepedia

- **PPIN is usually a set of PPIs; it is not put into biological context**

NUS National University of Singapore

Sources of Biological Pathways

Database	Remarks
KEGG	KEGG (`http://www.genome.jp/kegg`) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.
WikiPathways	WikiPathways (`http://www.wikipathways.org`) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.
Reactome	Reactome (`http::/www.reactome.org`) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.
Pathway Commons	Pathway Commons (`http://www.pathwaycommons.com`) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.
PathwayAPI	PathwayAPI (`http://www.pathwayapi.com`) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh *et al.*, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.

Source: Goh et al. “How advancement in biological network analysis methods empowers proteomics”. *Proteomics*, accepted.

Sources of Protein Interactions

Database	# nodes, # edges	URL	Build Focus	Reference
BioGRID	10k, 40k	http://thebiogrid.org	Literature	(Stark *et al.*, 2006)
DIP	2.6k, 3.3k	http://dip.doe-mbi.ucla.edu	Literature	(Xenarios *et al.*, 2002)
HPRD	30k, 40k	http://www.hprd.org	Literature	(Prasad *et al.*, 2009)
IntAct	56k, 267k	http://www.ebi.ac.uk/intact	Literature	(Aranda *et al.*, 2010)
MINT	30k, 90k	http://mint.bio.uniroma2.it/mint	Literature	(Chatr-aryamontri *et al.*, 2007)
STRING	5200k, ?	http://string-db.org	Literature, Prediction	(Szklarczyk *et al.*, 2011)

Source: Goh et al. "How advancement in biological network analysis methods empowers proteomics". *Proteomics*, accepted.

and Protein Complexes

- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, *NAR*, 2010

Use of Biological Networks in Enhancing Bioinformatics Analysis

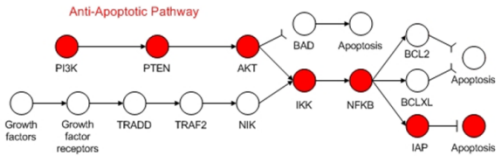
Limsoon Wong

Sivachenko et al. *JBCB*, 5(2b):429-546, 2007

Recall from Unit2 of the course…

Gene Expression Profile Analysis

Gene Regulatory Circuits



- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype

- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype

Contextualization!

Taming false positives by considering pathways instead of all possible groups

Group of Genes

- Suppose
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- What is the chance of a group of 5 genes being perfectly correlated to these samples?

- Prob(group of genes correlated) = $(1/2^6)^5$
  - Good, << $1/2^6$
- ~~# of groups = $^{100000}C_5$~~
- ~~E(# of groups of genes correlated) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$~~

# of pathways = 1000

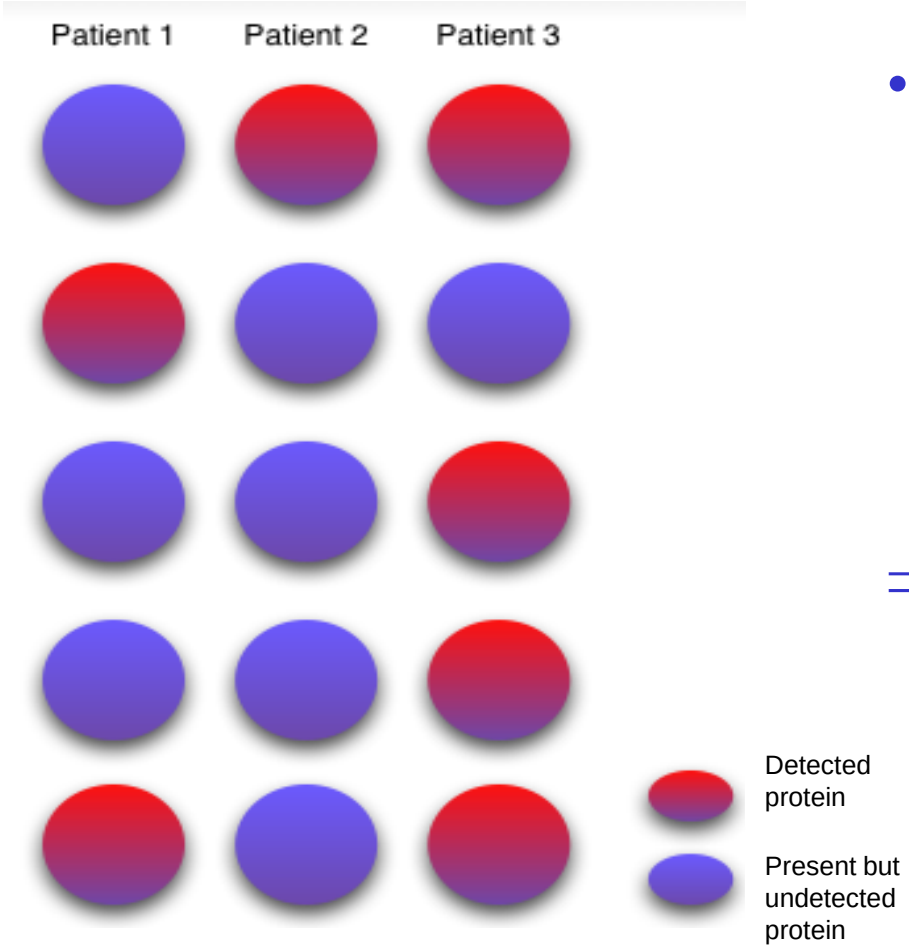
E(# of pathways correlated) = $1000 * (1/2^6)^5 = 9.3*10^{-7}$

⇒ Even more false positives?
- Perhaps no need to consider every group

Microarray Workshop for Gene Expression Profiling, NUH, 23/9/2011 Copyright 2011 © Limsoon Wong

Recall from Unit2 of the course…

Proteomic Profile Analysis



- **Suppose the failure to form a protein complex causes a disease**
  - If any component protein is missing, the complex can't form

⇒ **Diff patients suffering from the disease can have a diff protein component missing**

  - Construct a profile based on complexes?

Goh et al. How advancement in biological network analysis methods empowers proteomics. *Proteomics*, in press

Emily et al. *European Journal of Human Genetics*, 17(10):1231-1240, 2009.

Epistatic Interaction Mining

- **GWAS have linked many SNPs to diseases, but many genetic risk factors still unaccounted for**
- **Proteins coded by genes interact in cell**

⇒ **Some SNPs affect the phenotype in combination with other SNPs; i.e., epistasis**

- **Exhaustive search for epistatic effects has to test many combinations (>100,000^2) of SNPs**
  - Hard to get statistical significance
  - Take long time to run on computers

⇒ **Use biological networks to narrow the search for two-locus epistasis**

Vanunu et al. *PLoS Computational Biology*, 6(1):e1000641, 2010

NUS National University of Singapore

Disease Causal Gene Prioritization

- **Genes causing the same or similar diseases tend to lie close to one another in PPIN**

- **Given disease Q. Look for proteins in PPIN interacting with many causal genes of diseases similar to Q**

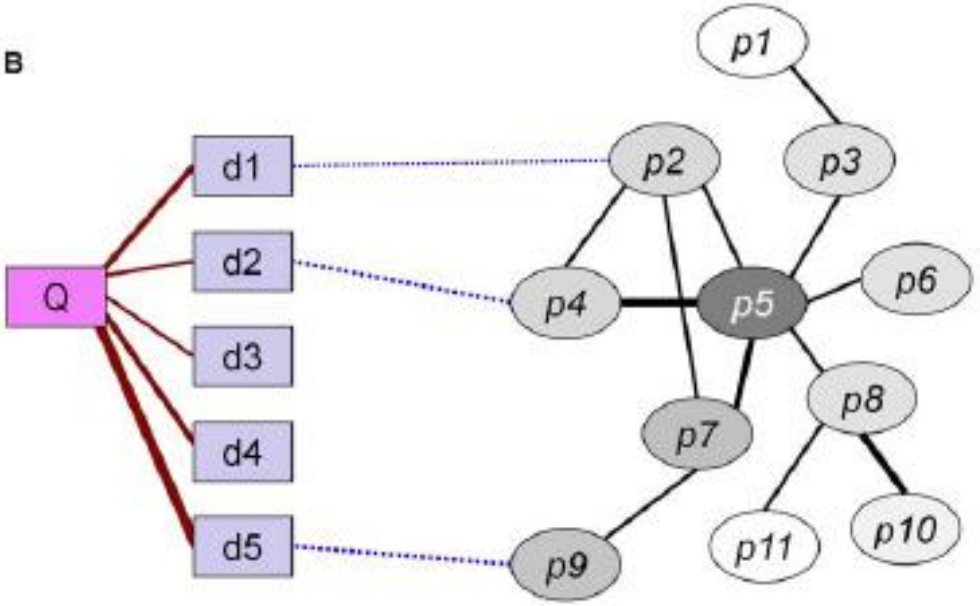
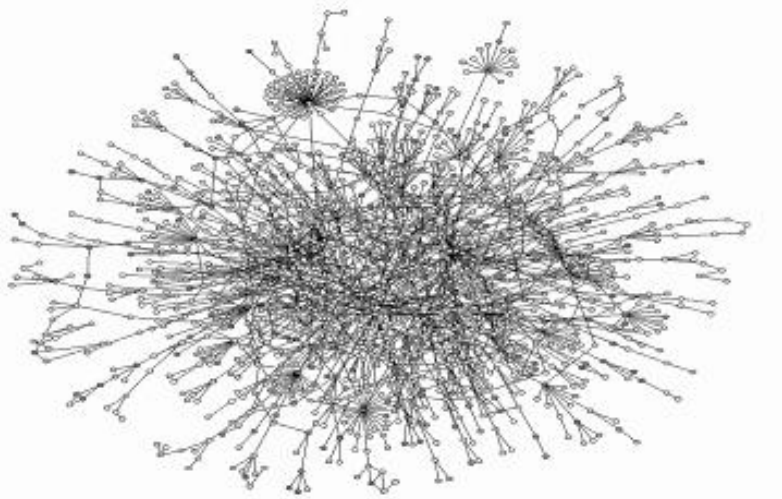


Figure 1. Illustration of the PRINCE algorithm. A query disease, denoted *Q*, has varying degrees of phenotypic similarity with other diseases, denoted *d1–d5* (marked with maroon lines, where thicker lines represent higher similarity). Known causal genes for these similar diseases are connected by dashed blue lines and used as the prior information. *p1–p11* comprise the protein set of a protein-protein interaction network, where interactions are marked with black lines and thicker lines denote edges with higher confidence. A scoring function that is smooth over the network is computed using an iterative network propagation method. At every iteration of the algorithm, each protein pumps flow to its neighbors and receives flow from them. Protein colors correspond to the flow they receive in a specific iteration, the darker the color the higher the flow. (A):

Liu et al. *Bioinformatics*, 25(15):1891-1897, 2009

Protein Complex Prediction

- **Nature of high-throughput PPI expts**
  - Proteins are taken out of their natural context!



- **Can a protein interact with so many proteins simultaneously?**

- **A big "hub" and its "spokes" should probably be decomposed into subclusters**
  - Each subcluster is a set proteins that interact in the same space &time; viz., a protein complex

- **Many complexes have highly connected cores in PPIN ➔ Find complexes by clustering**
- **Issue: How to identify low edge density complexes?**

Wong. *Internet Mathematics*, 7(4):274--298, 2011

Protein Function Prediction

- **Proteins with similar function are topologically close in PPIN**
  - Direct functional association
  - Indirect functional association

A pair of proteins that participate in the same cellular processes or localize to the same cellular compartment are many times more likely to interact than a random pair of proteins

- **Proteins with similar function have interaction neighborhoods that are similar**

When proteins in the neighborhood of a protein X have similar functions to proteins in the neighborhood of a protein Y, then proteins X & Y likely operate in similar environment

Consistency, Comprehensiveness, and Compatibility of Biological Pathway Databases

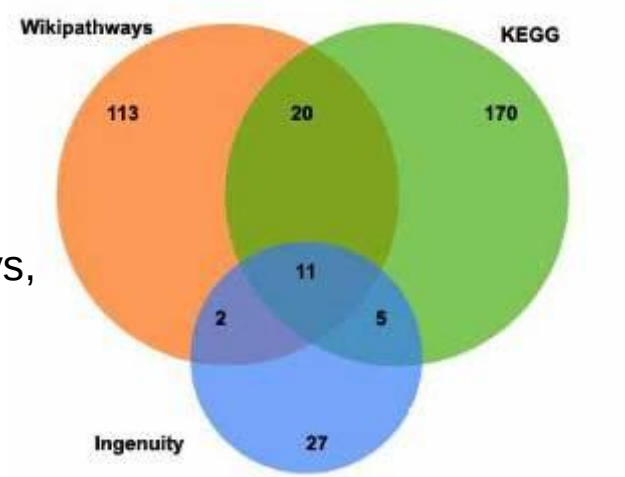
NUS National University of Singapore

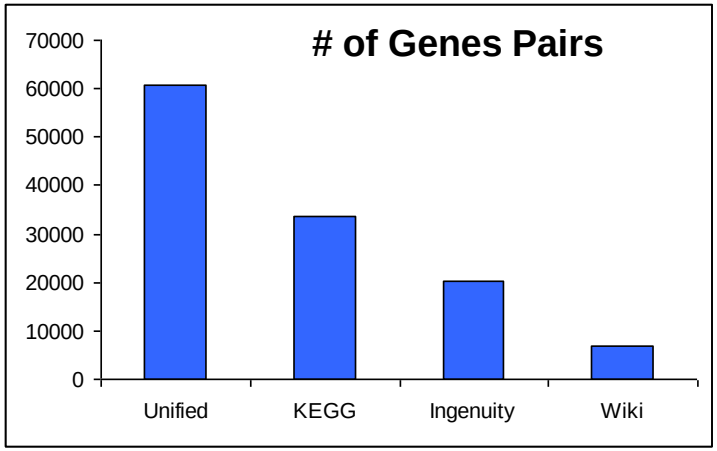
Major Sources of Biological Pathways

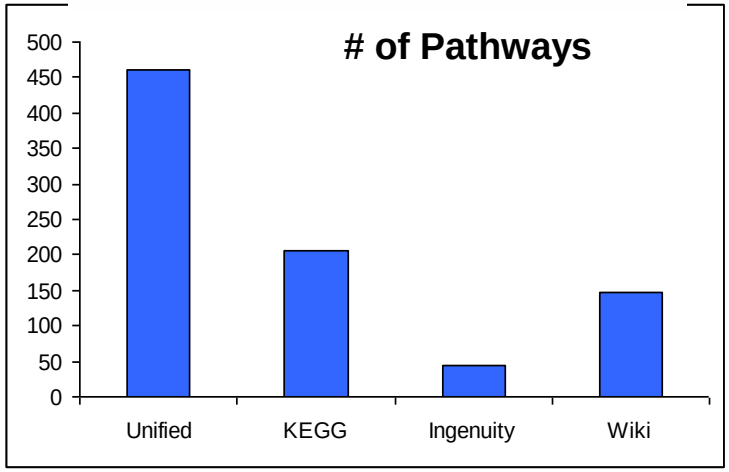
Database	Remarks
KEGG	KEGG (`http://www.genome.jp/kegg`) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.
WikiPathways	WikiPathways (`http://www.wikipathways.org`) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.
Reactome	Reactome (`http::://www.reactome.org`) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.
Pathway Commons	Pathway Commons (`http://www.pathwaycommons.com`) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.
PathwayAPI	PathwayAPI (`http://www.pathwayapi.com`) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh *et al.*, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.

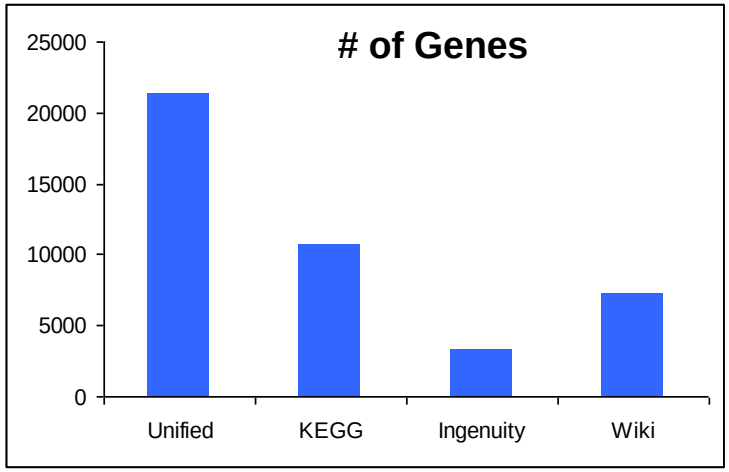
Low Comprehensiveness of Human Pathway Sources

Human pathways in Wikipathways, KEGG, & Ingenuity





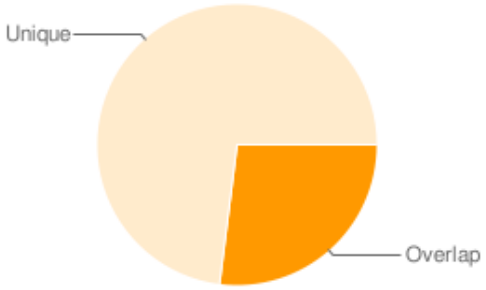




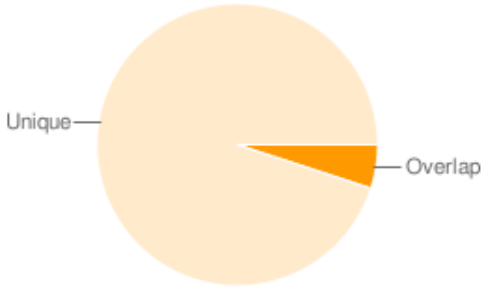
Soh et al. Consistency, Comprehensiveness, and Compatibility of Pathway Databases. *BMC Bioinformatics*, 11:449, 2010.

Low Consistency of Human Pathway Sources

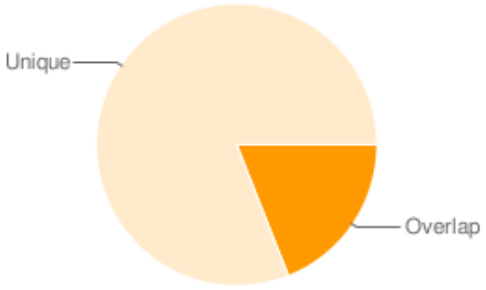
Gene Pair Overlap



Wiki vs KEGG

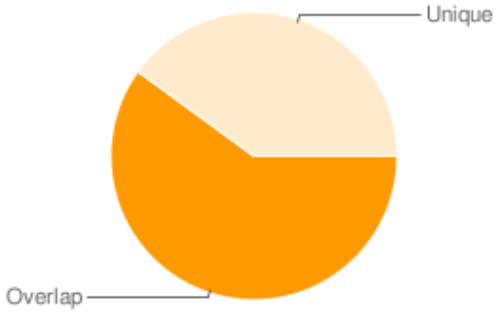


Wiki vs Ingenuity

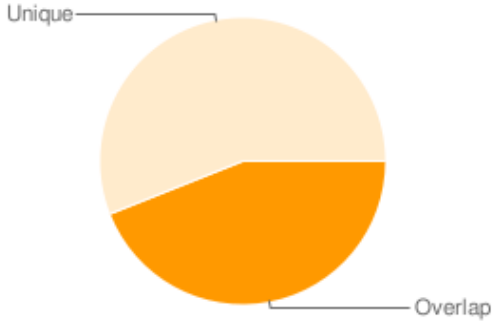


KEGG vs Ingenuity

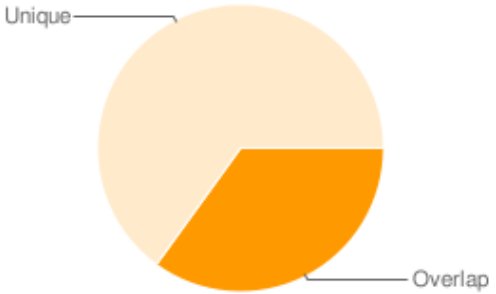
Gene Overlap



Wiki vs KEGG



Wiki vs Ingenuity



KEGG vs Ingenuity

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

Example: Human Apoptosis Pathway

Apoptosis Pathway			
	Wiki x KEGG	Wiki x Ingenuity	KEGG x Ingenuity
Gene Pair Count:	144 vs 172	144 vs 3557	172 vs 3557
Gene Count:	85 vs 80	85 vs 176	80 vs 176
Gene Overlap:	38	28	30
Gene % Overlap:	48%	33%	38%
Gene Pair Overlap:	23	14	24
Gene Pair % Overlap:	16%	10%	14%

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

The same low inter-database consistency (in gene overlap) is observed in pathways of other organisms

M. musculus	KEGG vs WikiPathways	WikiPathways vs MouseCyc	MouseCyc vs KEGG
Overlap Genes	2,611	532	919
Unique Genes	5,168	4,214	5,662
Jaccard Coefficient	0.336	0.112	0.140
S. cerevisiae	KEGG vs WikiPathways	WikiPathways vs YeastCyc	YeastCyc vs KEGG
Overlap Genes	801	402	480
Unique Genes	996	601	1,317
Jaccard Coefficient	0.446	0.400	0.267
M. tuberculosis H37Rv	KEGG vs WikiPathways	WikiPathways vs MTBRvCyc	MTBRvCyc vs KEGG
Overlap Genes	141	60	432
Unique Genes	948	525	707
Jaccard Coefficient	0.129	0.103	0.379

Source: Zhou Hufeng

The same low inter-database consistency (in gene pair overlap) is observed in pathways of other organisms

M. musculus	KEGG vs WikiPathways	WikiPathways vs MouseCyc	MouseCyc vs KEGG
Overlap Gene Pairs	875	1,242	2,068
Unique Gene Pairs	55,489	33,312	38,891
Jaccard Coefficient	0.016	0.036	0.050
S. cerevisiae	KEGG vs WikiPathways	WikiPathways vs YeastCyc	YeastCyc vs KEGG
Overlap Gene Pairs	35	9	419
Unique Gene Pairs	2,909	1,479	3,524
Jaccard Coefficient	0.012	0.006	0.106
M. tuberculosis H37Rv	KEGG vs WikiPathways	WikiPathways vs MTBRvCyc	MTBRvCyc vs KEGG
Overlap Gene Pairs	9	8	358
Unique Gene Pairs	3,819	2,810	5,823
Jaccard Coefficient	0.002	0.003	0.058

Source: Zhou Hufeng

Source: Zhou Hufeng

NUS National University of Singapore

Example: TCA Cycle Pathway

M. musculus	TCA cycle pathway	KEGG vs WikiPathways	KEGG vs MouseCyc	MouseCyc vs WikiPathways
Gene	Count	31 vs 30	31 vs 13	13 vs 30
	Overlap	24	13	11
	Jaccard Coefficient	0.65	0.42	0.34
Gene Pair	Count	100 vs 30	100 vs 24	24 vs 30
	Overlap	10	9	7
	Jaccard Coefficient	0.083	0.078	0.149

H. sapiens	Fatty Acid Biosynthesis	KEGG vs WikiPathways	KEGG vs HumanCyc	HumanCyc vs WikiPathways
Gene	Count	6 vs 22	6 vs 2	2 vs 22
	Overlap	3	2	1
	Jaccard Coefficient	0.12	0.33	0.04
Gene Pair	Count	12 vs 29	12 vs 2	2 vs 29
	Overlap	1	1	0
	Jaccard Coefficient	0.025	0.077	0.0

M. tuberculosis H37Rv	TCA cycle pathway	KEGG vs WikiPathways	KEGG vs MTBRvCyc	MTBRvCyc vs WikiPathways
Gene	Count	35 vs 34	35 vs 10	10 vs 34
	Overlap	34	10	10
	Jaccard Coefficient	0.97	0.29	0.29
Gene Pair	Count	107 vs 37	107 vs 19	19 vs 37
	Overlap	3	9	5
	Jaccard Coefficient	0.021	0.077	0.098

Incompatibility Issues

Pathway sources are curated. They are incomplete; but they have few errors. ➔ Makes sense to combine them. But…

- **Data extraction method variations**
- **Format variations**
- **Data differences**
- **Gene/GeneID name differences**
- **Pathway name differences**

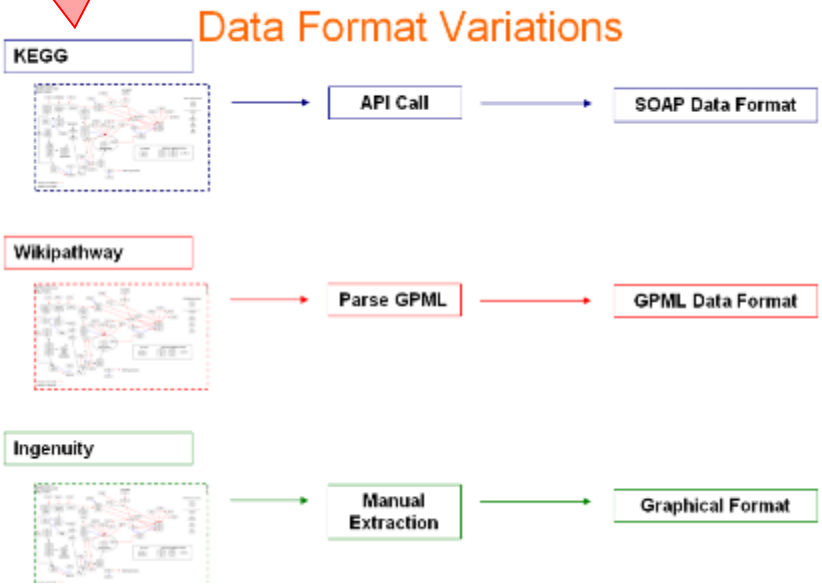


Image credit: Donny Soh's PhD dissertation, 2009

Integrating Pathway Databases

Things to deal with

- **Any integration of incompatible pathway databases must deal with**
  - Data extraction method variations
  - Format variations
  - Data differences
  - Gene name / gene id differences
  - Pathway name differences
- **We discuss only pathway name differences**
- **For other issues, consult**
  - Zhou et al. IntPath---an integrated pathway gene relationship database for model organisms and important pathogens, *BMC Bioinformatics*, submitted

The same pathways in the different sources are often given different names.

So how do we even know two pathways are the same and should be compared / merged?

Example of Pathway Name Differences

IntPath	KEGG	WikiPathways	MouseCyc
Fatty Acid Biosynthesis	Fatty acid biosynthesis	Fatty Acid Biosynthesis	1. fatty acid biosynthesis initiation II 2. very long chain fatty acid biosynthesis 3. fatty acid biosynthesis initiation III
Cholesterol Biosynthesis		Cholesterol Biosynthesis	1. cholesterol biosynthesis III (via desmosterol) 2. cholesterol biosynthesis II (via 24,25-dihydrolanosterol) 3. cholesterol biosynthesis I 4. superpathway of cholesterol biosynthesis
TCA cycle	Citrate cycle (TCA cycle)	TCA cycle	TCA Cycle
Glycolysis and Gluconeogenesis	Glycolysis / Gluconeogenesis	Glycolysis and Gluconeogenesis	1. glycolysis I 2. glycolysis II

Source: Zhou Hufeng

Possible Ways to Match Pathways

- **Match based on name (LCS)**
  - Pathways w/ similar name should be the same pathway
  - But annotations are very noisy

  ⇒Likely to mismatch pathways?

  ⇒Likely to match too many pathways?

- **Are the followings good alternative approaches?**
  - Match based on overlap of genes
  - Match based on overlap of gene pairs

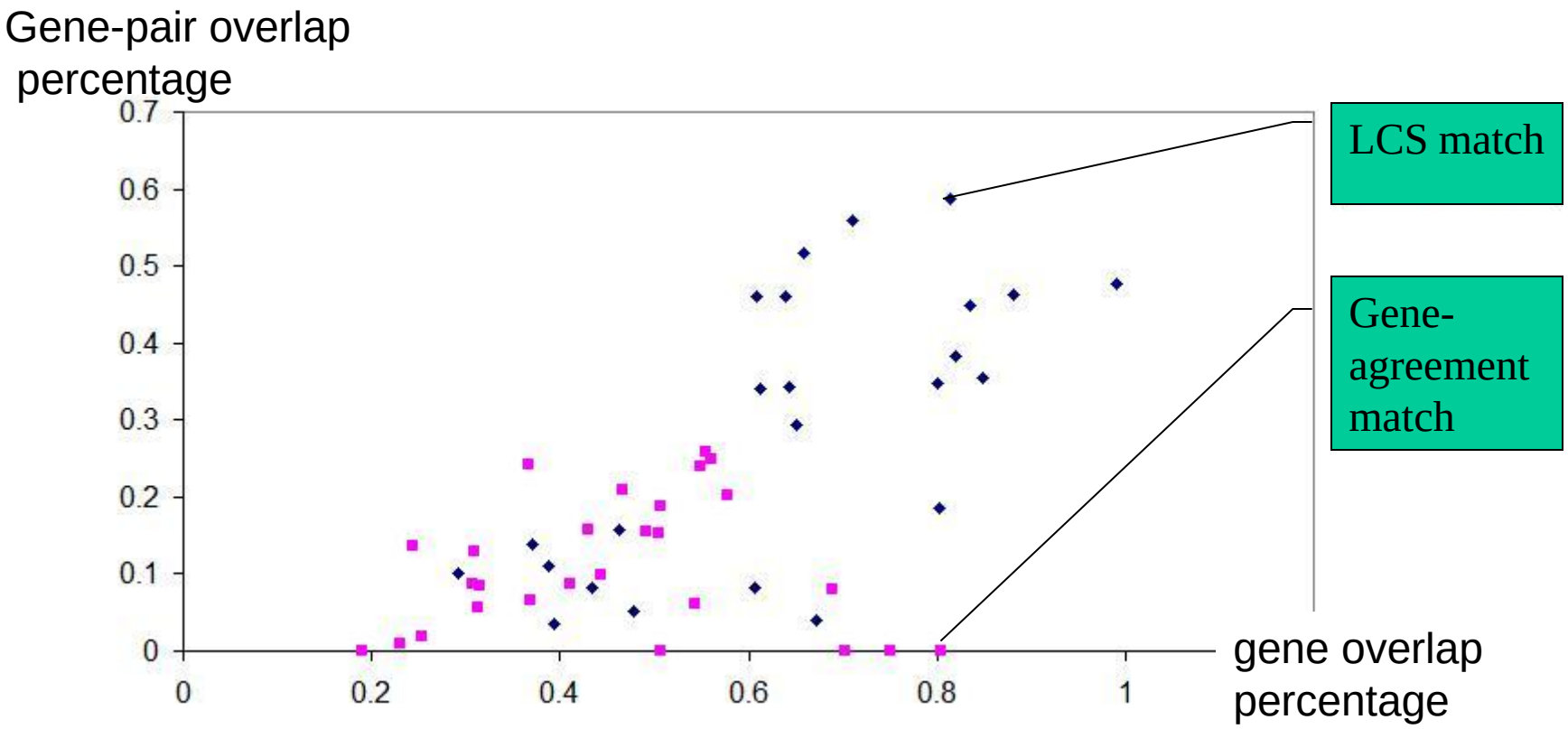
LCS vs Gene-Agreement Matching

- **Accuracy**
  - 94% of LCS matches are in top 3 gene agreement matches
  - 6% of LCS matches not in top 3 of gene agreement matches; but their gene-pair agreement levels are higher
- **Completeness**
  - Let Pi be pathway in db A that LCS cannot find match in db B
  - Let Qi be pathway in db B with highest gene agreement to Pi
  - Gene-pair agreement of Pi-Qi is much lower than pathway pairs matched by LCS

LCS is better than gene-agreement based matching!

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

LCS vs Gene-Agreement Matching

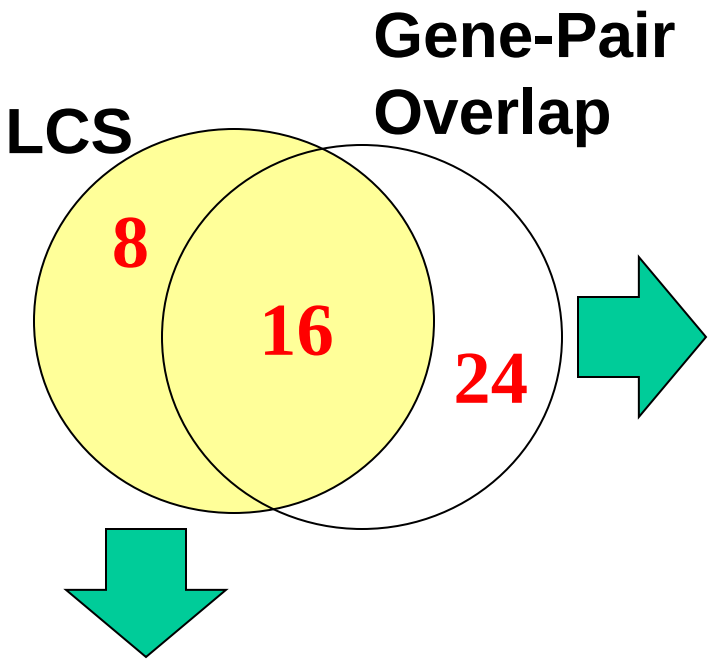


- **LCS consistently has higher gene-pair agreement**

$\Rightarrow$ **LCS is better than gene-agreement based matching!**

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

LCS vs Gene-Pair Agreement Matching



ErbB signaling pathway	JAK/Stat Signaling
Calcium signaling pathway	Synaptic Long Term Potentiation
Apoptosis	Toll-like receptor signaling pathway
VEGF signaling pathway	Axonal Guidance Signaling
Gap junction	PPAR-alpha/RXR-alpha Signaling
Natural killer cell mediated cytotoxicity	Fc Epsilon RI Signaling
T cell receptor signaling pathway	Axonal Guidance Signaling
B cell receptor signaling pathway	Axonal Guidance Signaling
Olfactory transduction	cAMP-mediated Signaling
GnRH signaling pathway	B Cell Receptor Signaling
Melanogenesis	Wnt Signaling Pathway and Pluripotency
Type II diabetes mellitus	Insulin Recpetor Signaling
Colorectal cancer	Toll-like receptor signaling pathway
Renal cell carcinoma	Axonal Guidance Signaling
Pancreatic cancer	PTEN Signaling
Endometrial cancer	PTEN Signaling
Glioma	ERK/MAPK Signaling
Prostate cancer	JAK/Stat Signaling
Basal cell carcinoma	Wnt Signaling Pathway and Pluripotency
Melanoma	FGF Signaling
Chronic myeloid leukemia	GM-CSF Signaling
Acute myeloid leukemia	PTEN Signaling
Small cell lung cancer	Toll-like receptor signaling pathway
Non-small cell lung cancer	GM-CSF Signaling

The 24 pathway pairs singled out by maximal gene-pair overlap

Regulation of actin cytoskeleton	Regulation of Actin Cytoskeleton
Wnt signaling pathway	Wnt Signaling Pathway
T cell receptor signaling	t cell receptor Signaling
VEGF signaling	VEGF Signaling
MAPK signaling	MAPK Cascade
Apoptosis	Apoptosis
Apoptosis	Apoptosis Signaling
Toll-like receptor	Toll-like receptor signaling pathway

The 8 pathway pairs singled out by LCS

Note: We consider only pathway pairs that have at least 20 reaction overlap.

LCS vs Gene-Pair Agreement Matching

- **Gene-pair agreement match will miss when**
  - Pathway P in db A has few overlap with pathway P in db B due to incompleteness of db, even if pathway name matches perfectly!
  - Example: wnt signaling pathway, VEGF signaling pathway, MAPK signaling pathway, etc. in KEGG don't have largest gene-pair overlap w/ corresponding pathways in Wikipathways & Ingenuity

$\Rightarrow$ **Bad for getting a more complete unified pathway P**

LCS vs Gene-Pair Agreement Matching

- **Pathways having large gene-pair overlap are not necessarily the same pathways**

- **Examples**
  - "Synaptic Long Term Potentiation" in Ingenuity vs "calcium signalling" in KEGG
  - "PPAR-alpha/RXR-alpha Signaling" in Ingenuity vs "TGF-beta signaling pathway" in KEGG

⇒ **Difficult to set correct gene-pair overlap threshold to balance against false positive matches**

Further Improvement to LCS

- **Please read the reference below for some of the improvements made to LCS**
  - Zhou et al. IntPath---an integrated pathway gene relationship database for model organisms and important pathogens, *BMC Bioinformatics*, submitted

An Interesting Question

- **If two pathways are merged, how do you choose the name of the resulting merged pathway?**
  - Pick the longer of the two original names?
  - Pick the shorter?
  - Pick randomly?
- **Why?**
- **Any exception?**

IntPath	KEGG	WikiPathways	MouseCyc
Fatty Acid Biosynthesis	Fatty acid biosynthesis	Fatty Acid Biosynthesis	1. fatty acid biosynthesis initiation II 2. very long chain fatty acid biosynthesis 3. fatty acid biosynthesis initiation III
Cholesterol Biosynthesis		Cholesterol Biosynthesis	1. cholesterol biosynthesis III (via desmosterol) 2. cholesterol biosynthesis II (via 24,25-dihydro-lanosterol) 3. cholesterol biosynthesis I 4. superpathway of cholesterol biosynthesis
TCA cycle	Citrate cycle (TCA cycle)	TCA cycle	TCA Cycle
Glycolysis and Gluco-neogenesis	Glycolysis / Gluconeoge-nesis	Glycolysis and Gluco-neogenesis	1. glycolysis I 2. glycolysis II

Source: Zhou Hufeng

What have we learned?

- **Significant lack of concordance betw db's**
  - Level of consistency for genes is 0% to 88%
  - Level of consistency for genes pairs is 0%-61%
  - Most db contains less than half of the pathways in other db's

- **Matching pathways by name is better than matching by gene overlap or gene-pair overlap**

Reliability of PPIN

Sources of Protein Interactions

Database	# nodes, # edges	URL	Build Focus	Reference
BioGRID	10k, 40k	http://thebiogrid.org	Literature	(Stark *et al.*, 2006)
DIP	2.6k, 3.3k	http://dip.doe-mbi.ucla.edu	Literature	(Xenarios *et al.*, 2002)
HPRD	30k, 40k	http://www.hprd.org	Literature	(Prasad *et al.*, 2009)
IntAct	56k, 267k	http://www.ebi.ac.uk/intact	Literature	(Aranda *et al.*, 2010)
MINT	30k, 90k	http://mint.bio.uniroma2.it/mint	Literature	(Chatr-aryamontri *et al.*, 2007)
STRING	5200k, ?	http://string-db.org	Literature, Prediction	(Szklarczyk *et al.*, 2011)

Source: Goh et al. "How advancement in biological network analysis methods empowers proteomics". *Proteomics*, accepted.

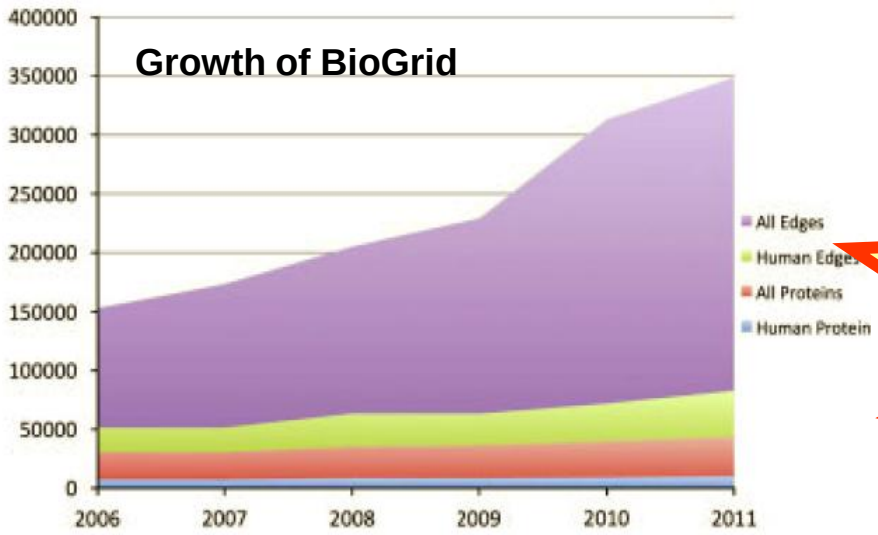
and Protein Complexes

- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, *NAR*, 2010

PPI Detection Assays

- **Many high-throughput assays for PPIs**
  - Y2H
  - TAP
  - Synthetic lethality

Generating *large amounts* of expt data on PPIs can be done with ease



- **But …**

High-throughput approaches sacrifice quality for ***quantity***: (a) limited or biased coverage: ***false negatives***, & (b) high error rates: ***false positives***

Noise in PPI Networks

Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results)	516	53	33
A2: GY2H Ito *et al.* (core)	798	64	40
A3: GY2H Ito *et al.* (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, *in vitro*	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

Sprinzak et al., *JMB*, 327:919-923, 2003

Large disagreement betw methods

- **High level of noise**

⇒ **Need to clean up before making inference on PPI networks**

Identifying Noise Edges in PPIN

Dealing with noise in PPIN using Reproducibility

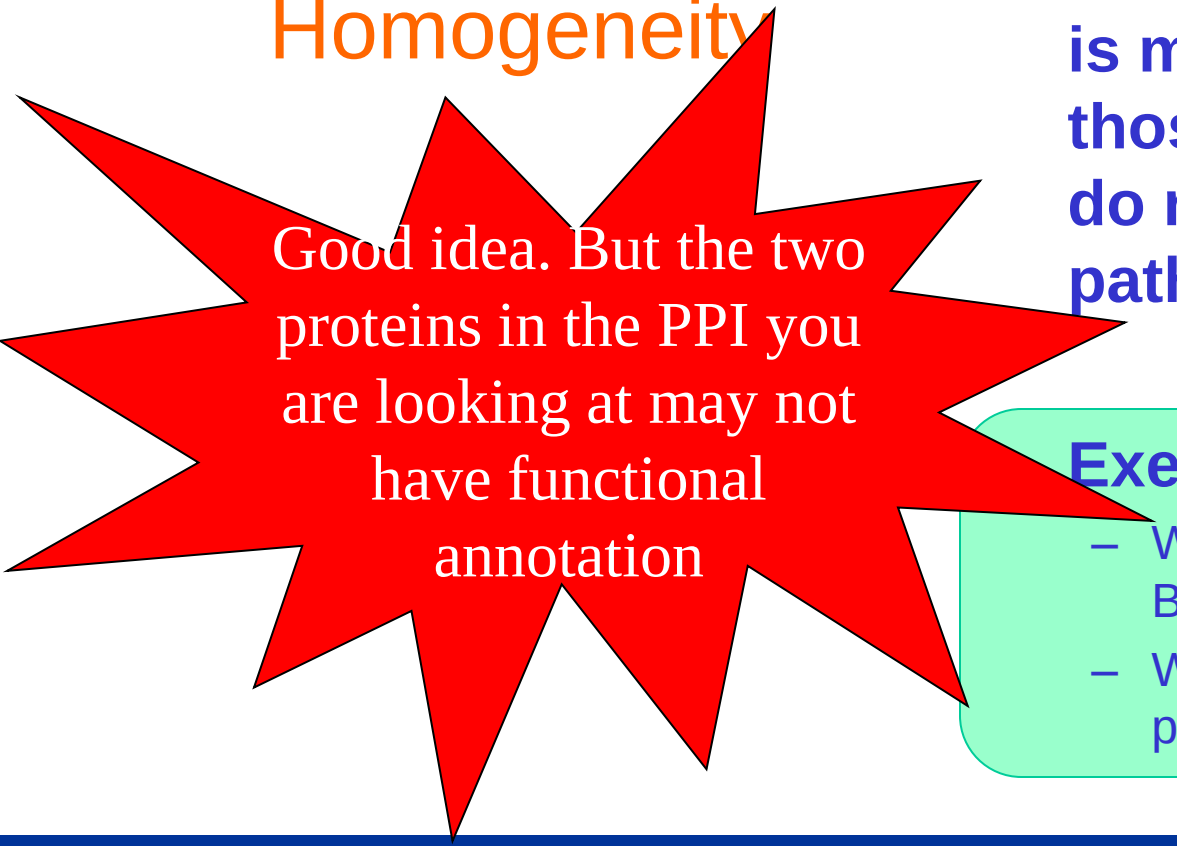
- **If a PPI is reported in a few independent expts, it is more reliable than those reported in only one expt**

$$r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i)$$

- r_i is reliability of expt source i,
- $E_{u,v}$ is the set of expt sources in which interaction betw u and v is observed

Good idea. But you need to do more expts
➔ More time & more $ has to be spent

Dealing with noise in PPIN using Functional Homogeneity



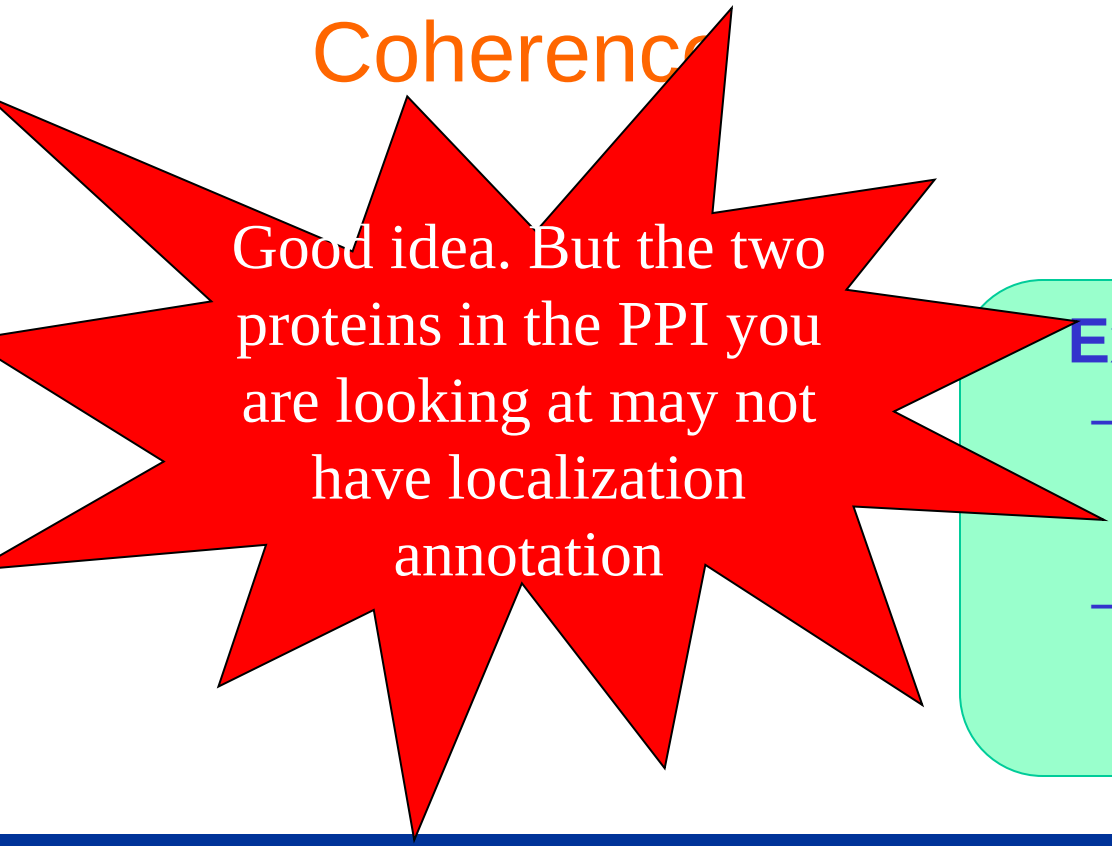
- **If two proteins in a PPI participate in the same function or pathway, it is more reliable than those whose proteins do not share function & pathway**

Exercise

- What fraction of yeast PPIs in BioGrid share function?
- What fraction of yeast protein pairs share function?

Dealing with noise in PPIN using Localization Coherence

- **Two proteins should be in the same place to interact. Agree?**



Exercise

- What fraction of yeast PPIs in BioGrid are in the same cellular compartment?
- What fraction of yeast protein pairs are in the same cellular compartment?

Chua & Wong. Increasing the Reliability of Protein Interactomes.
Drug Discovery Today, 13(15/16):652--658, 2008

Dealing with noise in PPIN using local topology around a PPI edge

- **Two proteins participating in same biological process are more likely to interact**
- **Two proteins in the same cellular compartments are more likely to interact**

- **CD-distance**
- **FS-Weight**

CD-distance & FS-Weight: Based on concept that two proteins with many interaction partners in common are likely to be in same biological process & localize to the same compartment

Brun, et al. *Genome Biology,* 5(1):R6, 2003

Czekanowski-Dice Distance

- **Given a pair of proteins (u, v) in a PPI network**
  - N_u = the set of neighbors of u
  - N_v = the set of neighbors of v
- **CD(u,v) =** $$\frac{2 \mid N_u \cap N_v \mid}{\mid N_u \mid + \mid N_v \mid}$$
- **Consider relative intersection size of the two neighbor sets, not absolute intersection size**
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, CD(u,v)=1
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, CD(u,v)=1

Liu et al. *GIW2008*, pp. 138-149

Iterated CD-Distance

- **Variant of CD-distance that penalizes proteins with few neighbors**

$$wL(u,v) = \frac{2\,|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

$$\lambda_u = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_u|\}, \lambda_v = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_v|\}$$

- **Suppose average degree is 4, then**
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, $wL(u,v) = 0.25$
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, $wL(u,v) = 1$

A thought…

$$wL(u,v) = \frac{2\,|\,N_u \cap N_v\,|}{|\,N_u\,| + \lambda_u + |\,N_v\,| + \lambda_v}$$

- **Weight of interaction reflects its reliability**

⇒ **Can we get better results if we use this weight to re-calculate the score of other interactions?**

Liu et al. *GIW2008*, pp. 138-149

Iterated CD-Distance

- $wL^0(u,v) = 1$ if $(u,v) \in G$, otherwise $wL^0(u,v)=0$
- $wL^1(u,v) = \dfrac{|N_u \cap N_v| + |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$
- $wL^k(u,v) = \dfrac{\sum_{x \in Nu \cap Nv} w_L^{k-1}(u,x) + \sum_{x \in Nu \cap Nv} w_L^{k-1}(v,x)}{\sum_{x \in Nu} w_L^{k-1}(u,x) + \lambda^k{}_u + \sum_{x \in Nv} w_L^{k-1}(v,x) + \lambda^k{}_v}$
- $\lambda^k{}_u = \max\{0, \dfrac{\sum_{x \in V} \sum_{y \in Nx} w_L^{k-1}(x,y)}{|V|} - \sum_{x \in Nu} w_L^{k-1}(u,x)\}$
- $\lambda^k{}_v = \max\{0, \dfrac{\sum_{x \in V} \sum_{y \in Nx} w_L^{k-1}(x,y)}{|V|} - \sum_{x \in Nv} w_L^{k-1}(v,x)\}$

Validation

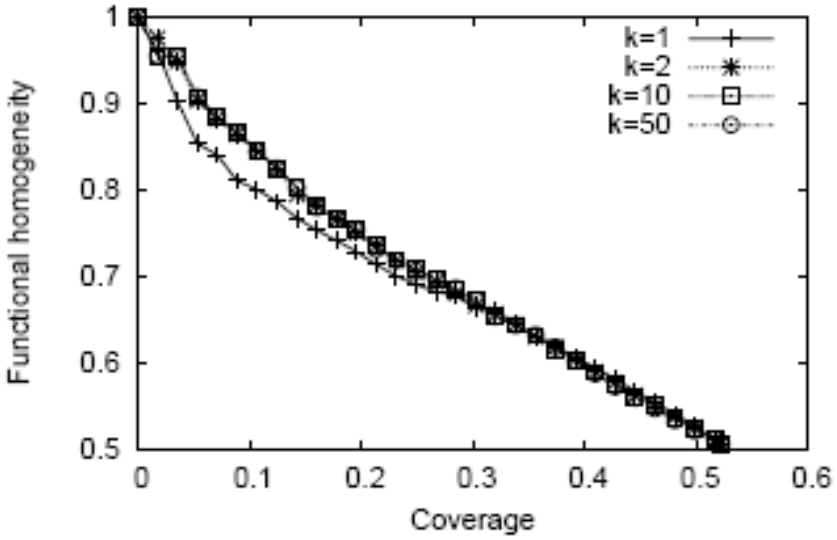
- **DIP yeast dataset**
  - Functional homogeneity is 32.6% for PPIs where both proteins have functional annotations and 3.4% over all possible PPIs
  - Localization coherence is 54.7% for PPIs where both proteins have localization annotations and 4.9% over all possible PPIs

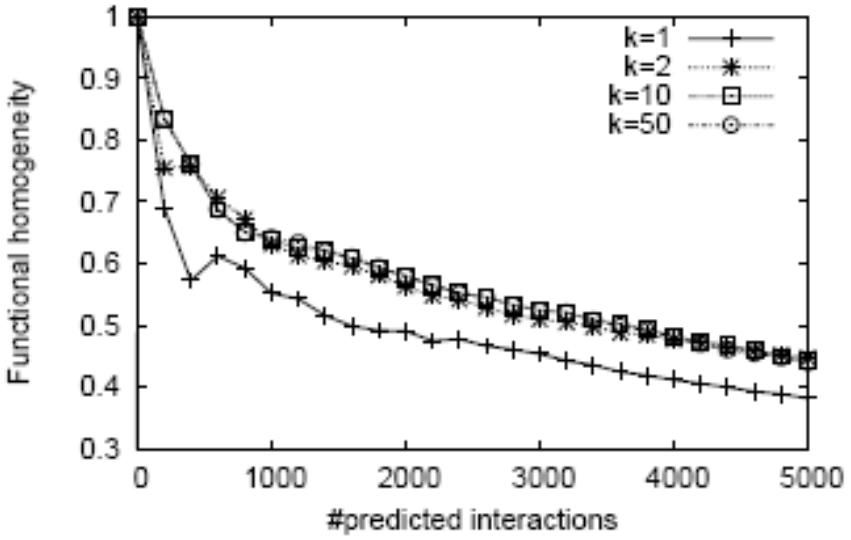
- **Let's see how much better iterated CD-distance is over the baseline above, as well as over the original CD-distance/FS-weight**

Liu et al. *GIW2008*, pp. 138-149

How many iteration is enough?

Cf. ave functional homogeneity of protein pairs in DIP < 4%
ave functional homogeneity of PPI in DIP < 33%





- **Iterated CD-distance achieves best performance wrt functional homogeneity at k=2**
- **Ditto wrt localization coherence (not shown)**

Liu et al. *GIW2008*, pp. 138-149

How many iteration is enough?

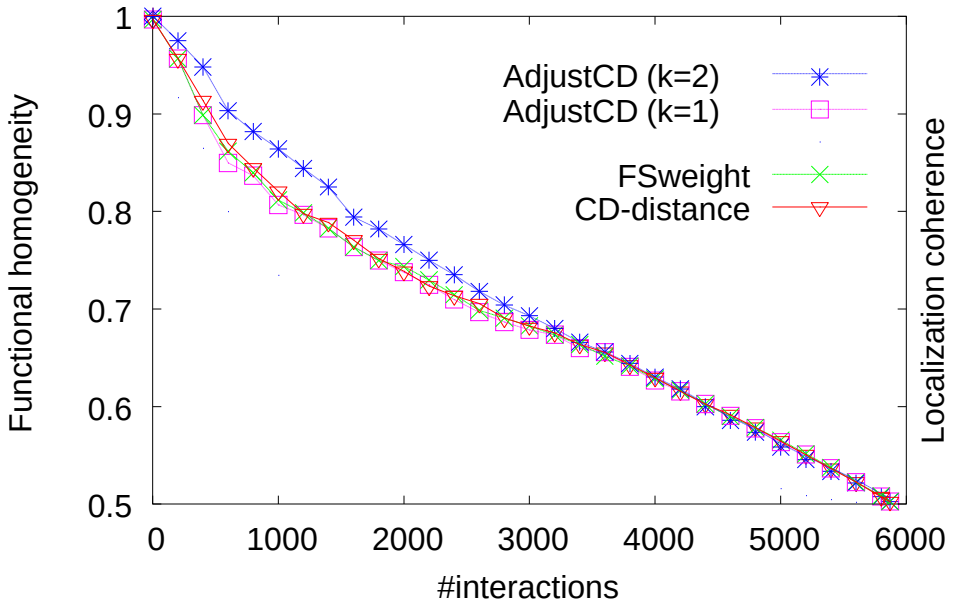
noise level	k	#common PPIs	avg_rank_diff	avg_score_diff
100%	1	5669	540.21	0.10
	2	5870	144.86	0.02
	20	5849	67.00	0.01
300%	1	5322	881.77	0.18
	2	5664	367.45	0.06
	20	5007	249.85	0.02
500%	1	5081	1013.14	0.23
	2	5502	625.46	0.12
	20	5008	317.33	0.05
1000%	k=1	4472	1187.10	0.28
	k=2	5101	1021.69	0.27
	k=20	5264	614.66	0.13

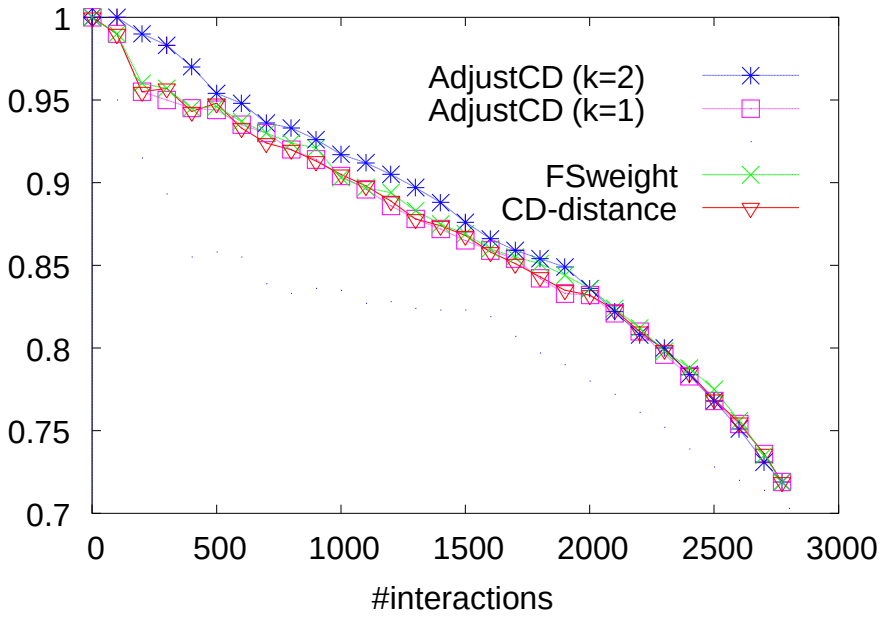
- **Iterative CD-distance at diff k values on noisy network**

$\Rightarrow$ # of iterations depends on amt of noise

Identifying False Positive PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%



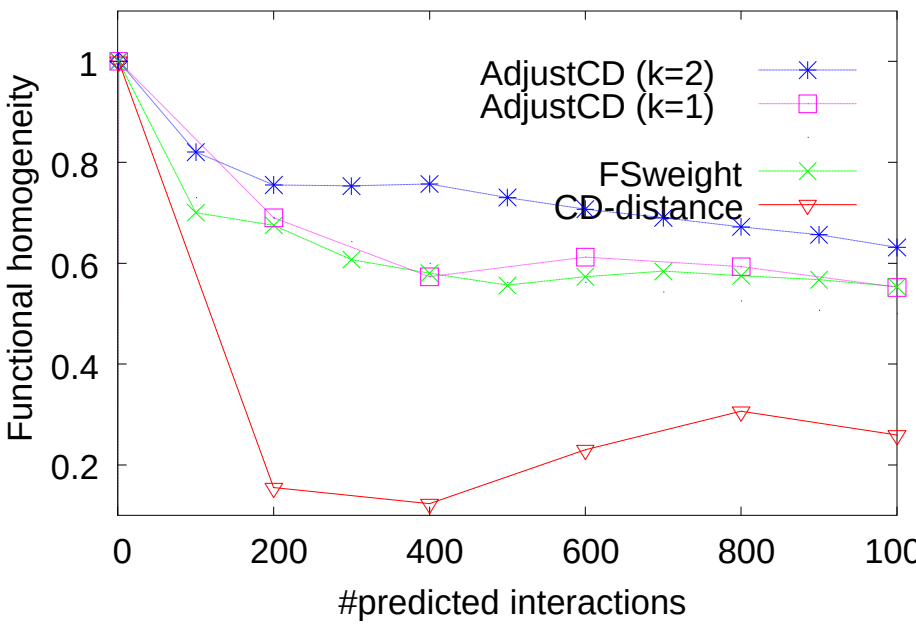


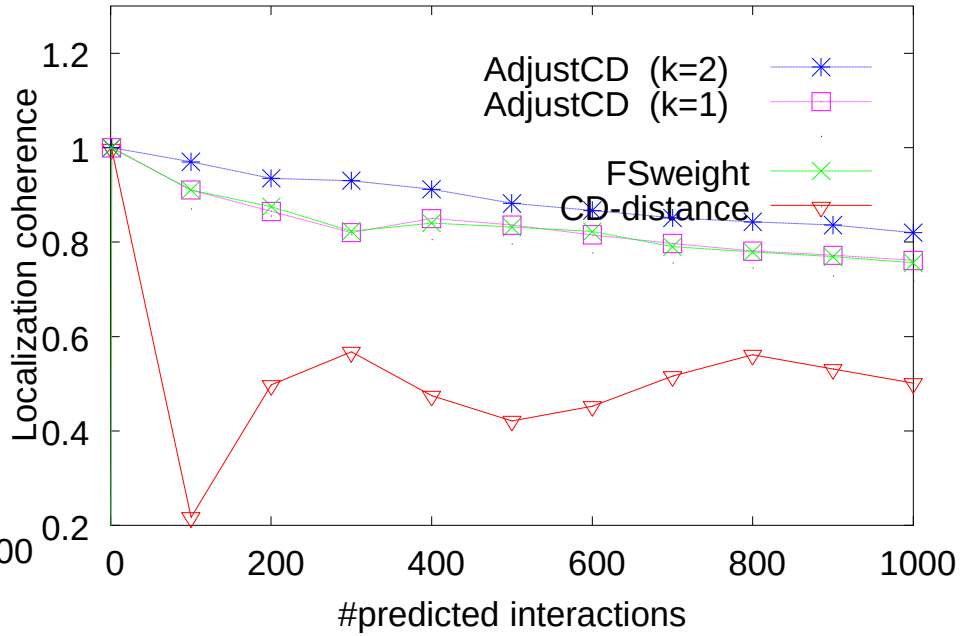
- **Iterated CD-distance is an improvement over previous measures for assessing PPI reliability**

Liu et al. *GIW2008,* pp. 138-149

Identifying False Negative PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%





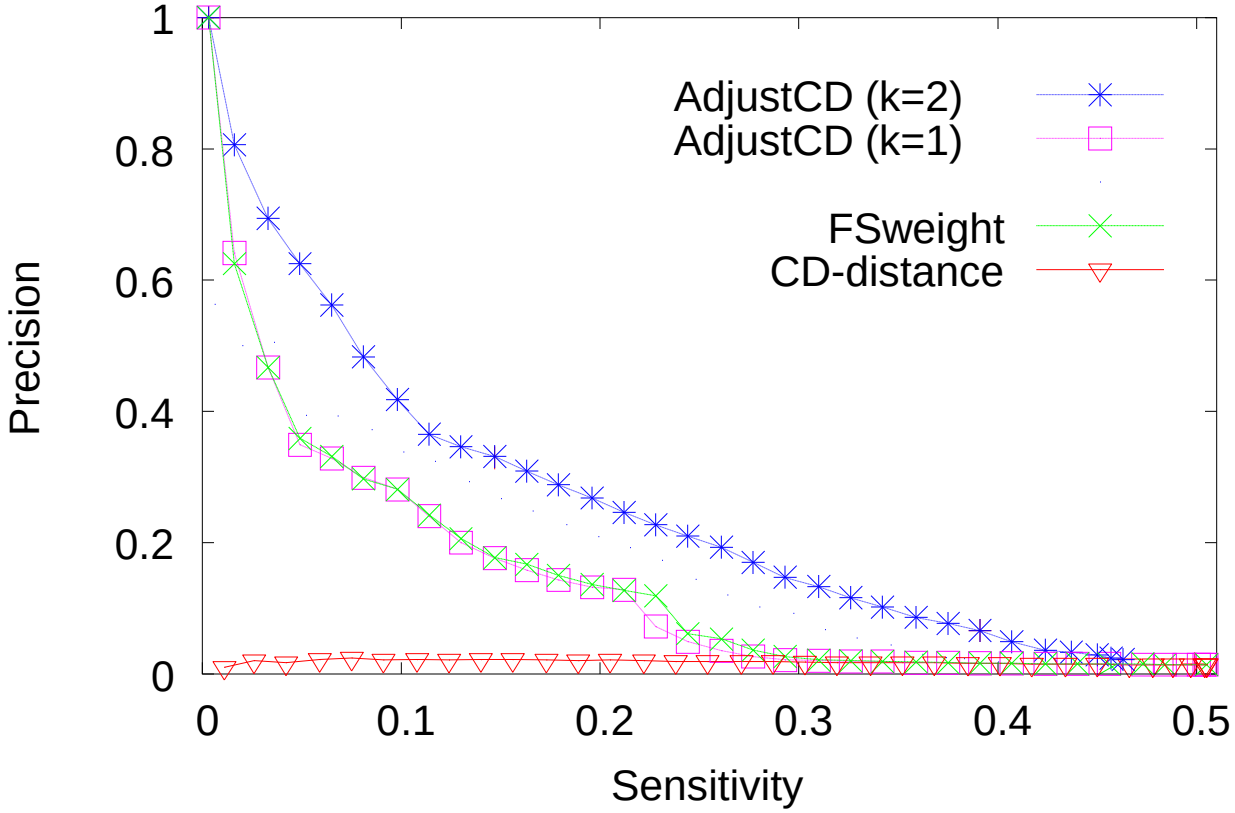
- **Iterated CD-distance is an improvement over previous measures for predicting new PPIs**

Liu et al. *GIW2008*, pp. 138-149

5-Fold Cross-Validation

- **DIP core dataset**
  - Ave # of proteins in 5 groups: 986
  - Ave # of interactions in 5 training datasets: 16723
  - Ave # of interactions in 5 testing datasets: 486591
  - Ave # of correct answer interactions: 307
- **Measures:**
  - sensitivity =TP/(TP + FN)
  - specificity =TN/(TN + FP)
    - **#negatives >> #positives, specificity is always high**
    - **>97.8% for all scoring methods**
  - precision =TP/(TP + FP)

5-Fold X-Validation



- **Iterated CD-distance is an improvement over previous measures for identifying false positive & false negative PPIs**

Liu et al. *GIW2008,* pp. 138-149

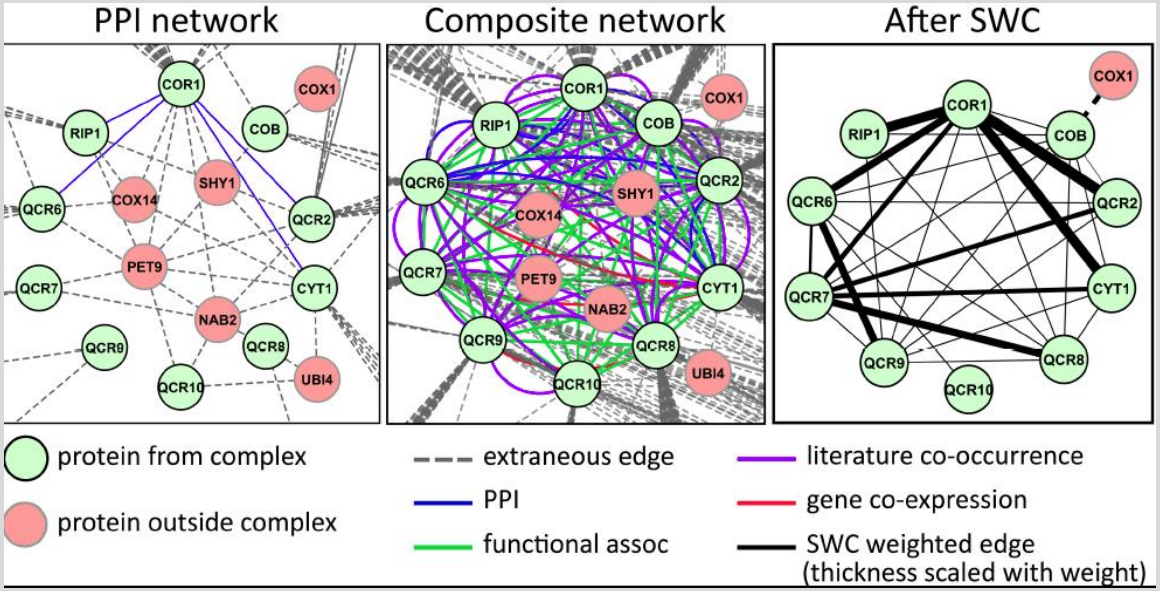
Combining multiple types of info to predict whether a PPI edge is real

- **Sometimes you do have additional independent info available**
  - Several PPI expts
  - Functional annotations
  - Localization information
- **You can combine these pieces of info in the following standard way:**

$$r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i)$$

  - r_i is reliability of expt source i,
  - $E_{u,v}$ is the set of expt sources in which interaction betw u and v is observed

Another way to combine more types of info to predict if a PPI is real



- **Overlay literature co-occurrence, gene co-expression, etc. on PPIN**
- **Machine learning to learn characteristic of real PPI**

$$
\begin{aligned}
& weight_{raw}(e) \\
= \; & P(e \; is \; comp | F_1 = f_1, F_2 = f_2, \ldots) \\
= \; & \frac{P(F_1 = f_1, F_2 = f_2, \ldots | e \; is \; comp) P(e \; is \; comp)}{P(F_1 = f_1, F_2 = f_2, \ldots)} \\
= \; & \frac{\prod_i P(F_i = f_i | e \; is \; comp) P(e \; is \; comp)}{\prod_i P(F_i = f_i)}
\end{aligned}
$$

Source: Yong Chern Han

Identifying Missing Edges in PPIN

Shoemaker & Panchenko. *PLoS Computational Biology*, 3(4):e43, 2007

NUS National University of Singapore

PPI Prediction Methods

Method Name	Protein/Domain Interaction	Physical Interaction/ Functional Association
Gene co-expression	P	F
Synthetic lethality	P	F
Gene cluster and gene neighbor	P	F
Phylogenetic profile	P, D	F
Rosetta Stone	P	F
Sequence co-evolution	P, D	F
Classification	P, D	P
Integrative	P, D	P
Domain association	D	P
Bayesian networks	P, D	F, P
Domain pair exclusion	D	P
p-Value	D	P

Second column shows if method is designed to predict protein (P) or domain (D) interactions (note that predicted domains can also be used for verifying protein interactions).
Third column shows if the method can be used to infer direct physical interaction (P) or indirect functional association (F).

You can also use our earlier topology scores, e.g, CD-distance to predict novel PPIs

Dandekar et al. *Trends Biochem Sci*, 23:324–328, 1998

PPI Prediction by Gene Clusters

- **Gene clusters or operons encoding co-regulated genes are usually conserved, despite shuffling effects of evolution**

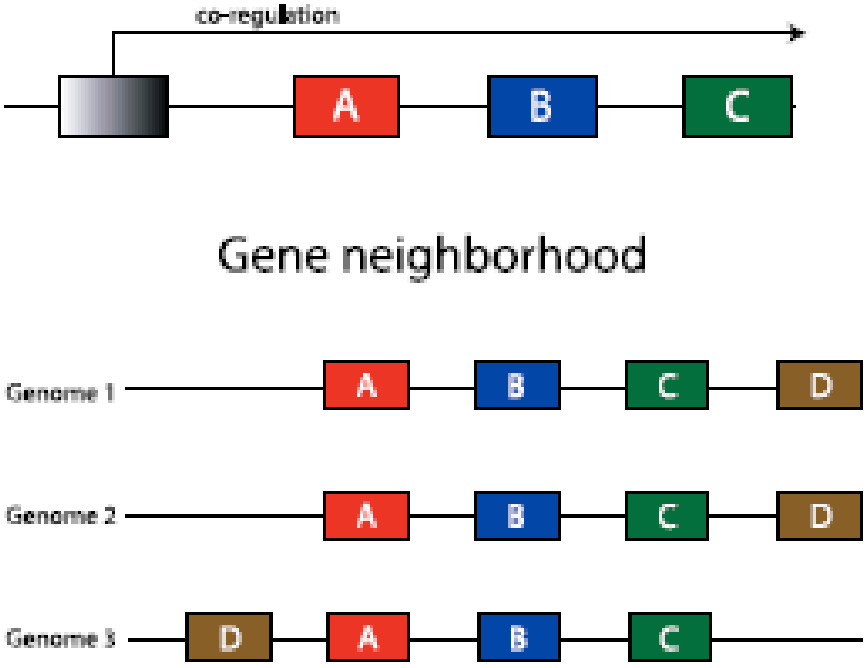


Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

⇒ Find conserved gene clusters

- Predict the genes to interact & form operons

Pellegrini et al. *PNAS*, 96:4285–4288, 1999

PPI Prediction by Phylogenetic Profiling

- **Components of complexes and pathways should be present simultaneously in order to perform their functions**
- Functionally linked and interacting proteins co-evolve and have orthologs in the same subset of fully sequenced organisms

	Genomes		
Proteins	EC	HI	BS
P1	0	1	1
P2	0	0	1
P3	1	0	0
P4	0	1	1

→ P1 and P4 are functionally linked

Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

Enright et al. *Nature,* 402:86–90, 1999

PPI Prediction by Rosetta Stone

- **Some interacting proteins have homologs in other genomes that are fused into one protein chain, a so-called Rosetta Stone protein**
- Gene fusion occurs to optimize co-expression of genes encoding for interacting proteins

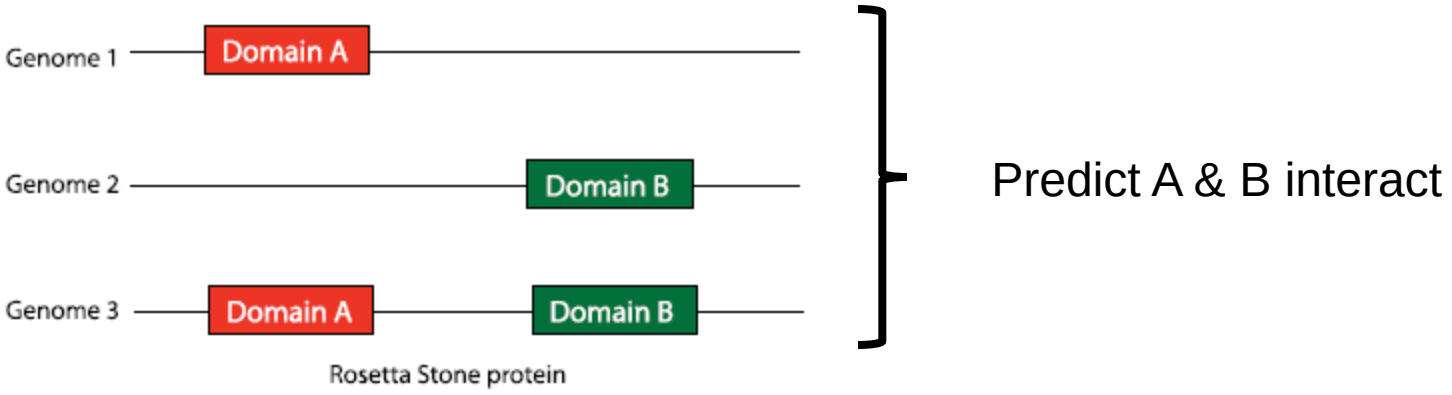


Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

See [Juan et al, *PNAS*, 105(3):934-939, 2008] for an impt further development to this idea

PPI Prediction by Seq Co-Evolution

- **Interacting proteins co-evolve**
  - Changes in one protein leading to loss of function are compensated by correlated changes in another protein
- Co-evolution is quantified by correlation of distance matrices used to construct the trees

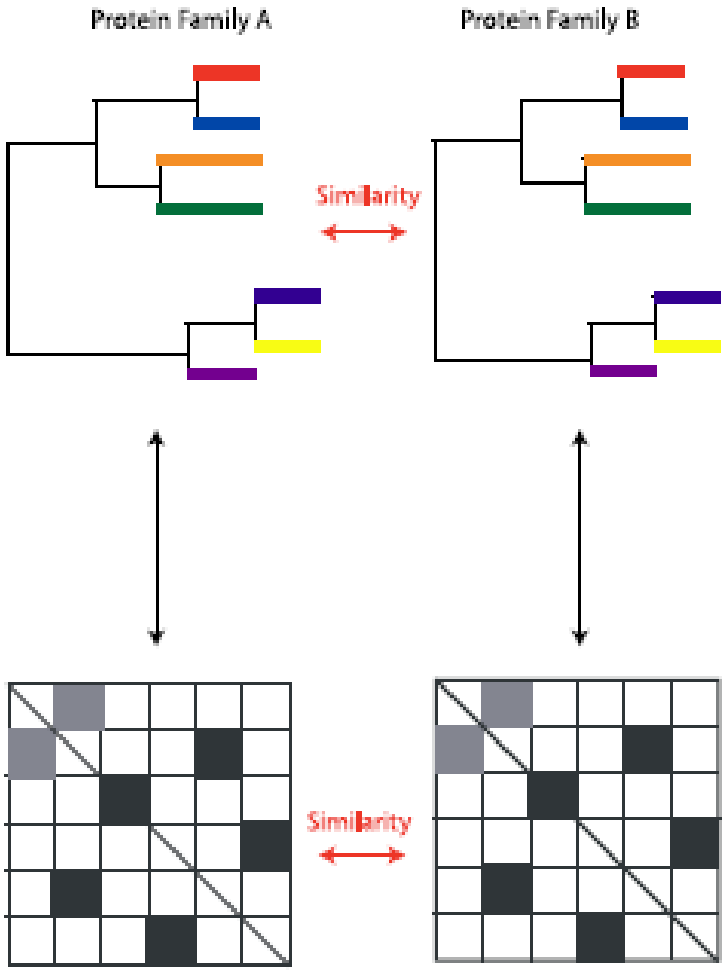
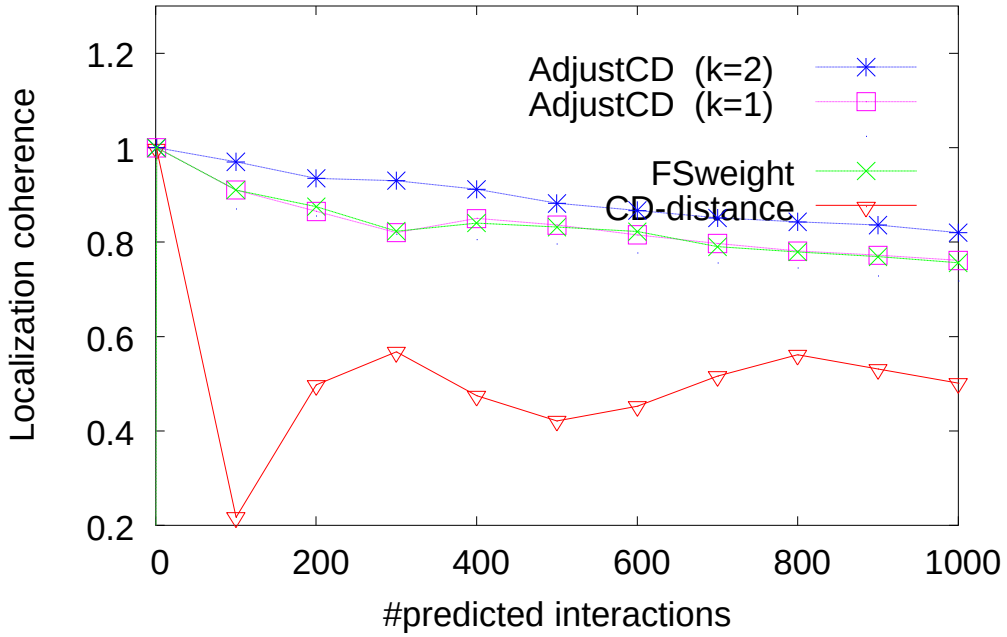


Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

Liu et al. *GIW2008*, pp. 138-149

PPI Prediction by Iterated CD-Distance

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%



$$wL^k(u,v) = \frac{\sum_{x \in Nu \cap Nv} wL^{k-1}(u,x) + \sum_{x \in Nu \cap Nv} wL^{k-1}(v,x)}{\sum_{x \in Nu} wL^{k-1}(u,x) + \lambda^k{}_u + \sum_{x \in Nv} wL^{k-1}(v,x) + \lambda^k{}_v}$$

- Predict (u,v) interact if $wL^k(u,v)$ is large

What have we learned?

- **It is possible to predict PPIs using a variety of information and methods**
  - Gene cluster, gene fusion, phylogenetic profile, sequence co-evolution, …

For those who are interested to go further:

- **How do you predict cross-species PPI's between a host and a pathogen?**

Must Read

- Zhou et al. **IntPath---an integrated pathway gene relationship database for model organisms and important pathogens,** *BMC Bioinformatics*, submitted
- Ng & Tan. **Discovering protein-protein interactions.** *JBCB*, 1(4):711-741, 2004
- Chua & Wong. **Increasing the Reliability of Protein Interactomes.** *Drug Discovery Today*, 13(15/16):652--658, 2008
- Shoemaker & Panchenko. **Deciphering protein-protein Interactions. Part II. Computational methods to predict protein and domain interaction partners.** *PLoS Computational Biology*, 3(4):e43, 2007

Good to Read

- Soh et al. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases.** *BMC Bioinformatics*, 11:449, 2010
- Sprinzak et al. **How reliable are experimental protein-protein interaction data?.** *JMB*, 327:919-923, 2003
- Liu et al. **Assessing and predicting protein interactions using both local and global network topological metrics.** *GIW 2008*, pp. 138-149
- Juan et al. **High-confidence prediction of global interactomes based on genome-wide coevolution networks.** *PNAS*, 105(3):934-939, 2008
- Enright et al. **Protein interaction maps for complete genomes based on gene fusion events.** *Nature,* 402:86–90, 1999
- Pellegrini et al. **Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles.** *PNAS*, 96:4285–4288, 1999
- Dandekar et al. **Conservation of gene order: A fingerprint of proteins that physically interact.** *Trends Biochem Sci*, 23:324–328, 1998

Comparative Analysis and Assessment of M. tuberculosis H37Rv Protein-Protein Interaction Datasets

Outline

- **Low similarity betw two MTB PPI datasets**
- **Hypothesis: One or both of them are of very poor quality?**

This part of the lecture is based on

- **Zhou & Wong. "Comparative Analysis and Assessment of *M. Tuberculosis* H37Rv Protein-Protein Interaction Datasets". *BMC Genomics*, 12(Suppl. 3):S20, 2011**

- **Evaluating the quality of the two datasets**
  - Informative GO assessment
  - PPI Functional Intensity matrix
  - Correlation w/ gene expression profiles
  - Interologs from diff organisms' expt PPI's
  - Integrated-pathway gene-pair relationships

Low Similarity of H37Rv PPI Datasets

- **STRING v 8.3 M. tuberculosis H37Rv PPI datasets**

STRING database prediction method	Number of PPIs	STRING database prediction method	Number of PPIs
Neighbourhood	12,706	Transferred neighborhood	78,376
Co-expression	0	Transferred co-expression	4,393
Experiments	4	Transferred experimental	4,129
Databases	7,030	Transferred databases	629
Text mining	2,715	Transferred text mining	11,074
Gene fusion	2,646	Co-occurrence	159,213
All STRING database PPIs	248,574		

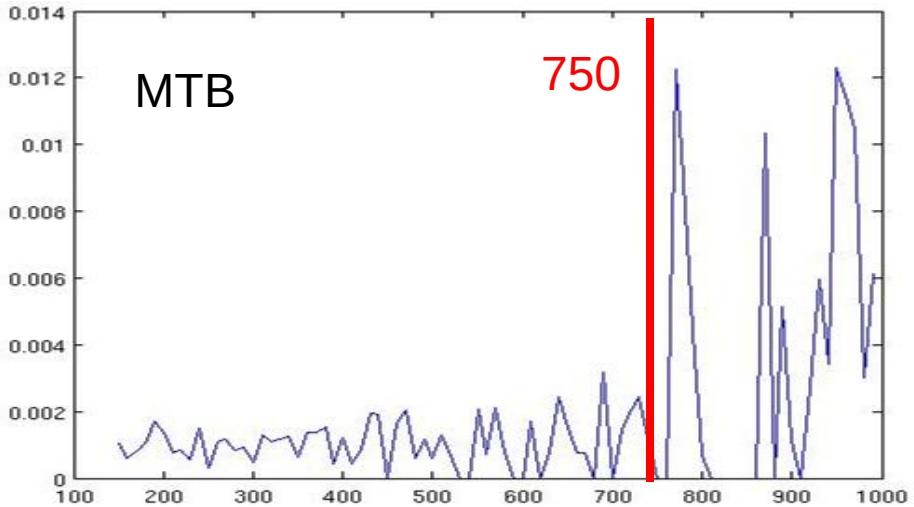
Summary of various subsets of H37Rv PPIs in STRING and their sources.

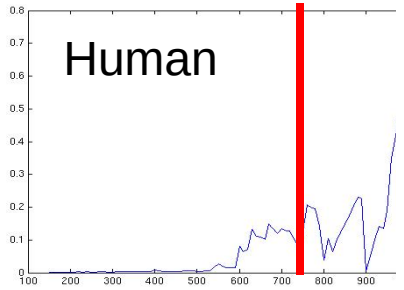
- **High-throughput bacterial two-hybrid (B2H) dataset: 8042 PPIs covering 2907 proteins**
  - Wang et al. “Global protein-protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv”. J Proteome Res, 9:6665-6677, 2010
- **Overlap # of PPIs betw the two datasets: 276**

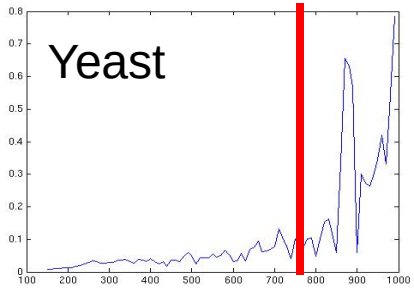
Why is the overlap so low?

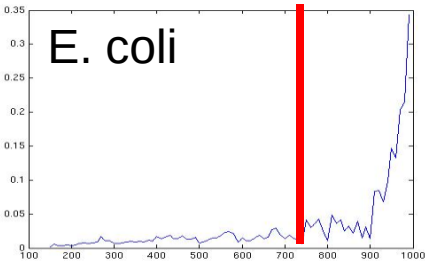
- **The STRING PPI dataset is too low quality?**
  - MTB PPI's in STRING come from predictions. Perhaps these predictions are all wrong
- **The MTB B2H PPI dataset is too low quality?**
  - B2H & Y2H technologies are very noisy. Perhaps these PPIs are produced by bad B2H expt
- **Both of the datasets are too low quality?**
- **Let us try to find out which of the above is the case**

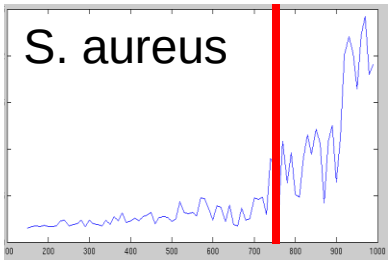
Fraction of Overlap Expt PPIs wrt STRING Score







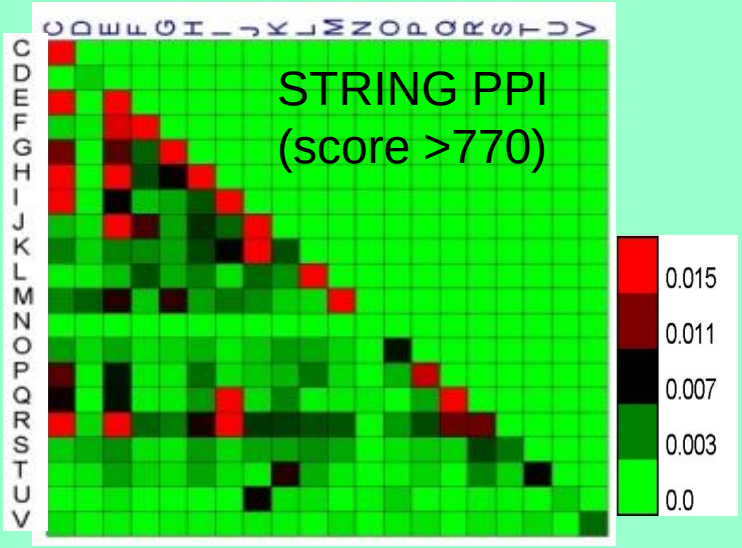




- **STRING and MTB B2H PPI dataset have higher similarity when STRING score is > ~750**
- **Similar observation for many other species**

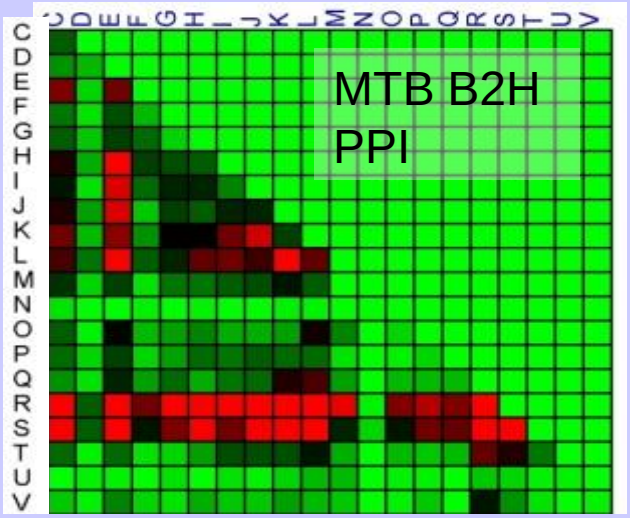
⇒ **Set threshold at ~750 for high-quality STRING PPI's**

Functional Intensity Matrix



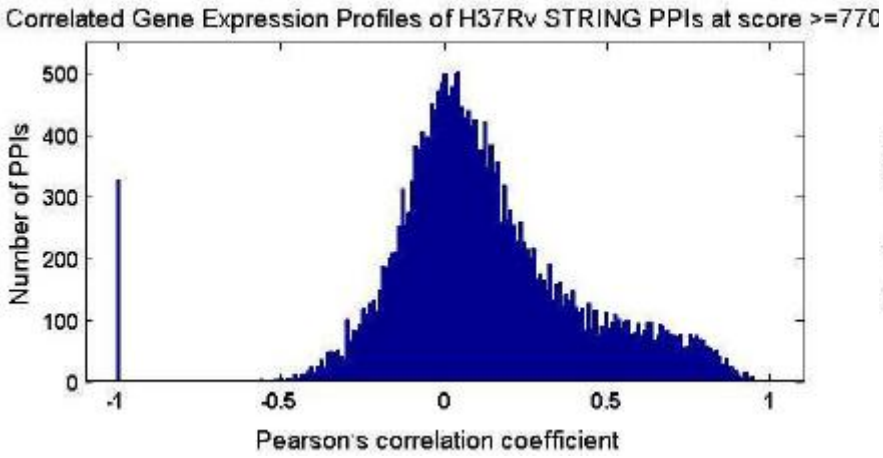
- **STRING PPI (score >770) shows strong diagonal intensity**

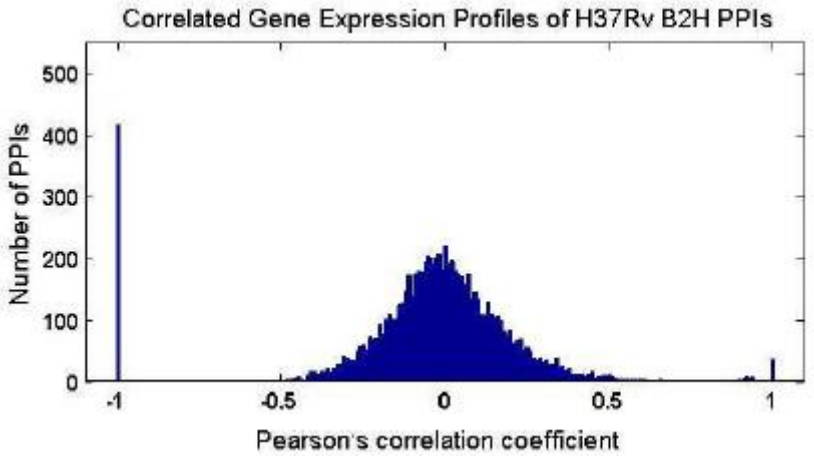
⇒ **Many PPIs have interacting partners with same function**



- **MTB B2H PPI dataset contains lots of PPI that are functional uninformative or unknown**

Correlated Gene Expression Profiles





- **Large portion of STRING PPI (score >770) have correlated gene expression profiles**
- **Many MTB B2H PPI's do not have correlated gene expression profiles in the two interacting partners**
- Note that PPIs with either of interaction partner w/o gene expression profile are assigned a Pearson's correlation coefficient to be -1.0. MTB B2H PPI dataset has large # of PPIs w/o gene expression profile

Informative GO Term Coherence

62% of STRING PPIs (score >770) have the same informative GO term in both proteins

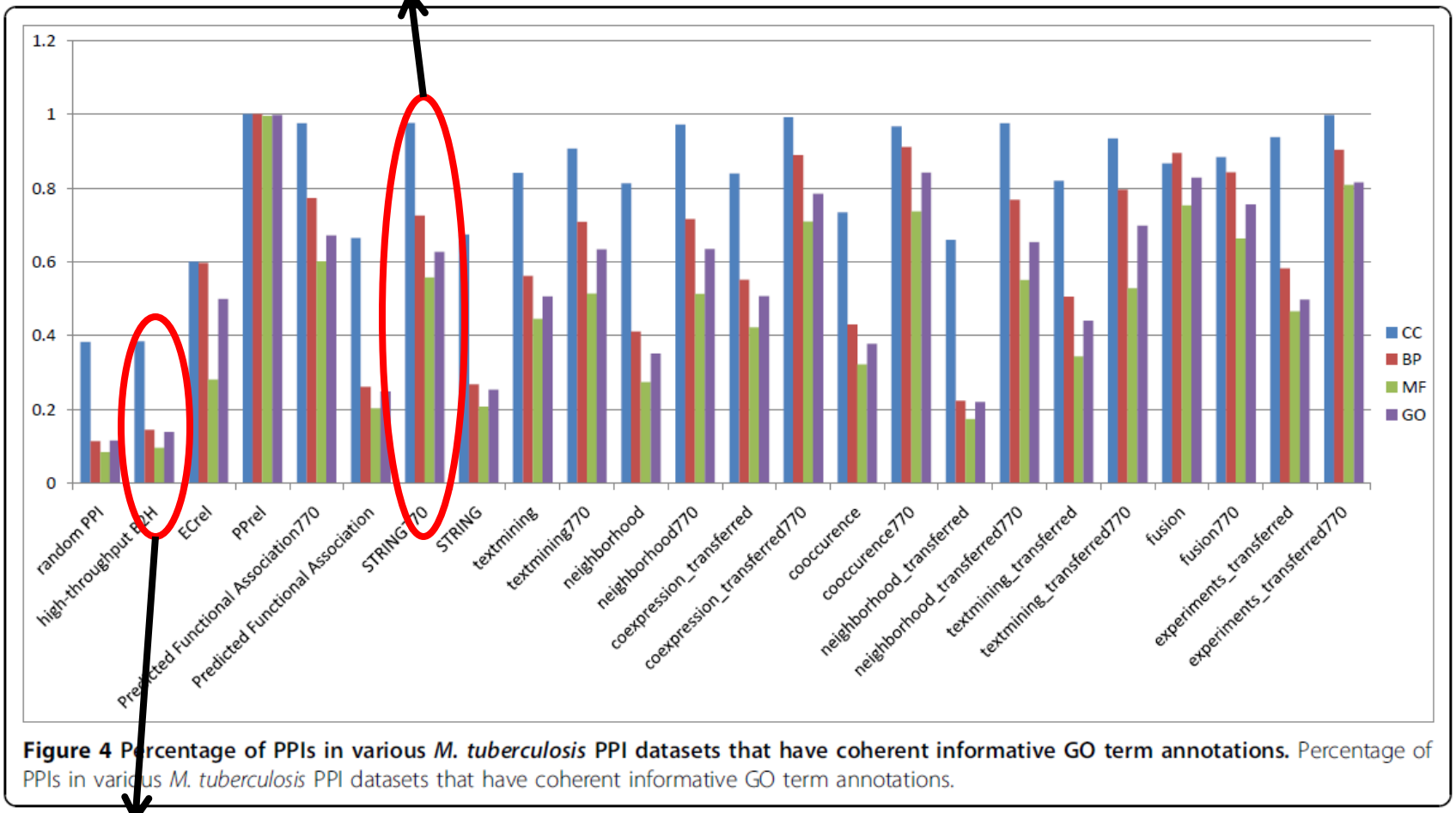


Figure 4 Percentage of PPIs in various *M. tuberculosis* PPI datasets that have coherent informative GO term annotations. Percentage of PPIs in various *M. tuberculosis* PPI datasets that have coherent informative GO term annotations.

Only 13% of MTB B2H PPI's show informative GO term coherence. No diff from random PPI's

GO Term Coherence: Comparison w/ Other Y2H PPI's

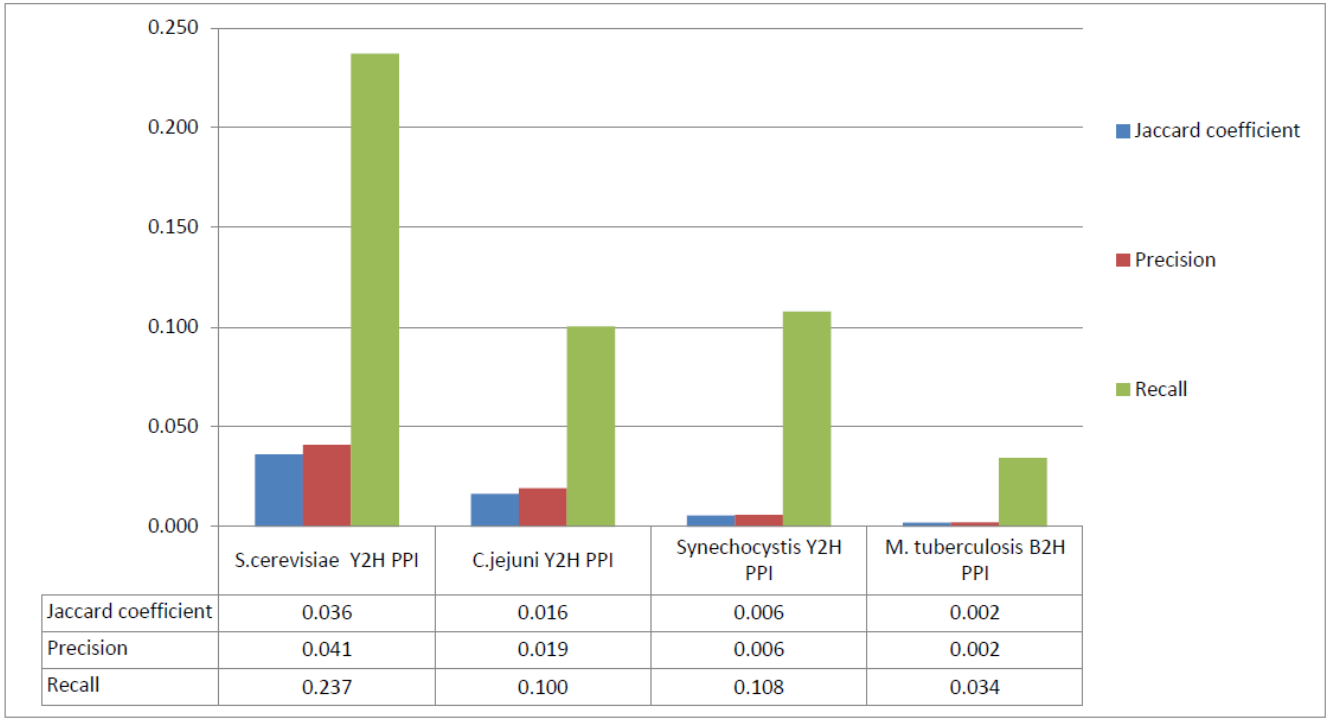
	CC	BP	MF	GO
S. cerevisiae Y2H PPI	49.74%	38.43%	35.90%	40.58%
S. cerevisiae random PPI	5.03%	3.70%	7.72%	7.08%
Info GO ratio of *S. cerevisiae* (Y2H PPI/ random)	9.88	10.38	4.65	5.73
Info GO term No.	57	173	78	365
C. jejuni Y2H PPI	47.37%	17.01%	16.27%	17.91%
C. jejuni random PPI	35.61%	11.04%	12.91%	13.97%
Info GO ratio of *C. jejuni* (Y2H PPI / random)	1.33	1.54	1.26	1.28
Info GO term No.	3	26	22	51
Synechocystis Y2H PPI	80.77%	26.16%	25.30%	29.47%
Synechocystis random PPI	44.94%	6.89%	8.59%	9.73%
Info GO ratio of *Synechocystis* (Y2H PPI / random)	1.80	3.79	2.95	3.03
Info GO term No.	3	30	30	63
M. tuberculosis B2H PPI	38.46%	14.36%	9.57%	13.83%
M. tuberculosis random PPI	38.30%	11.35%	8.35%	11.51%
Info GO ratio of *M. tuberculosis* (B2H PPI / random)	1.00	1.26	1.15	1.20
Info GO term No.	3	32	32	67

The table is a summary of the percentage of PPIs having coherent informative GO terms, number of informative GO terms and "Info GO ratio" in each of the representative two-hybrid PPI datasets in four organisms.

Info GO ratio = percentage of PPIs in two-hybrid PPI dataset having coherent informative GO terms / percentage of PPIs in random PPI dataset having coherent informative GO terms.

- **MTB B2H PPI has the lowest GO term coherence among all Y2H/B2H datasets, after normalization**

STRING PPI Overlap: Comparison w/ Other Y2H PPI's



	S.cerevisiae Y2H PPI	C.jejuni Y2H PPI	Synechocystis Y2H PPI	M. tuberculosis B2H PPI
Jaccard coefficient	0.036	0.016	0.006	0.002
Precision	0.041	0.019	0.006	0.002
Recall	0.237	0.100	0.108	0.034

Figure 7 Similarity of representative two-hybrid PPI datasets to corresponding STRING predicted functional associations datasets. The histogram is a summary of calculated Jaccard coefficient, precision and recall between a representative two-hybrid PPI dataset (as benchmark) and its corresponding STRING predicted functional associations dataset (as testing dataset).

- **MTB B2H PPI has much lower overlap with STRING PPI compared to other Y2H PPI datasets**

How well do MTB PPIs in STRING & B2H agree with interologs predicted from expt PPIs of other species

Table 3 Results of predicted interologs from STRING experimental PPIs

Source PPIs for homology transfer	STRING database experimental PPIs		
Identify homology and transfer	PIDE >30 Coverage >0.2 E-VALUE< 1*e-10	HVAL >20 E-VALUE <1*e-6	Orthologues Identified by Inparanoid (default parameters)
Benchmark	*M.tuberculosis* H37Rv high-throughput B2H PPIs dataset		
Jaccard coefficient	0.00354	0.00187	0.00289
Precision	0.00375	0.00196	0.00339
Recall	0.0588	0.0384	0.0190
Overlapping PPIs	5	4	5
Benchmark	*M.tuberculosis* H37Rv predicted functional associations dataset from STRING database(with PPIs score above 770)		
Jaccard coefficient	0.285	0.179	0.129
Precision	0.355	0.236	0.239
Recall	0.625	0.422	0.219
Overlapping PPIs	656	830	525

Predicted interologs from STRING experimental PPIs and a summary of Jaccard coefficient, precision and recall between the interologs datasets and two benchmarks.

- **Interologs predicted from expt PPIs of related species agree better w/ MTB PPIs in STRING**

Conclusion

Why is the overlap so low?

- ~~**The STRING PPI dataset is too low quality?**~~
  - ~~MTB PPI's in STRING come from predictions. Perhaps these predictions are all wrong~~
- **The MTB B2H PPI dataset is too low quality?**
  - B2H & Y2H technologies are very noisy. Perhaps these PPIs are produced by bad B2H expt
- ~~**Both of the datasets are too low quality?**~~
- ~~**Let us try to find out which of the above is the case**~~

Acknowledgements

Lui Guimei

Donny Soh

Zhou Hufeng

- **Lots of ideas on PPIN cleansing here came from my colleague ,Dr Liu Guimei**
- **Lots of results on consistency & comprehensiveness of pathway databases, and their integration, came from my students, Donny Soh & Zhou Hufeng**
- **The slides on MTB B2H PPI quality assessment are from the work of my student, Zhou Hufeng**