Increasing Confidence of Protein-Protein Interactomes

Limsoon Wong

(Based on work of/with Jin Chen, Kenny Chua, Wynne Hsu, Mong Li Lee, See-Kiong Ng, Rintaro Saito, Wing-Kin Sung)

Outline

- **Reliability of experimental protein-protein interaction data**
- **Identification of false positives**
  - Interaction generality
  - Interaction generality 2
  - Interaction pathway reliability
  - FS Weight
  - Meso-scale network motifs

How Reliable are Experimental Protein-Protein Interaction Data?

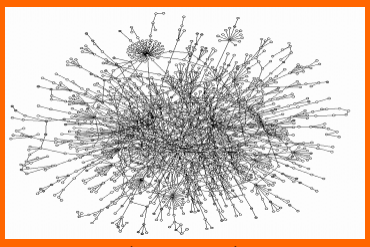


Figure credit: Jeong et al. 2001

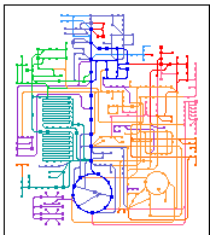
Why Protein Interactions?

- 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

High-Tech Expt PPI Detection Methods

- **Yeast two-hybrid assays**
- **Mass spec of purified complexes (e.g., TAP)**
- **Correlated mRNA expression**
- **Genetic interactions (e.g., synthetic lethality)**
- **...**

FACT: Generating ***large amounts*** of experimental data about protein-protein interactions can be done with ease.

Slide credit: See-Kiong Ng

Key Bottleneck

- **Many high-throughput expt detection methods for protein-protein interactions have been devised**
- **But ...**

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

Slide credit: See-Kiong Ng

Some Protein Interaction Data Sets

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

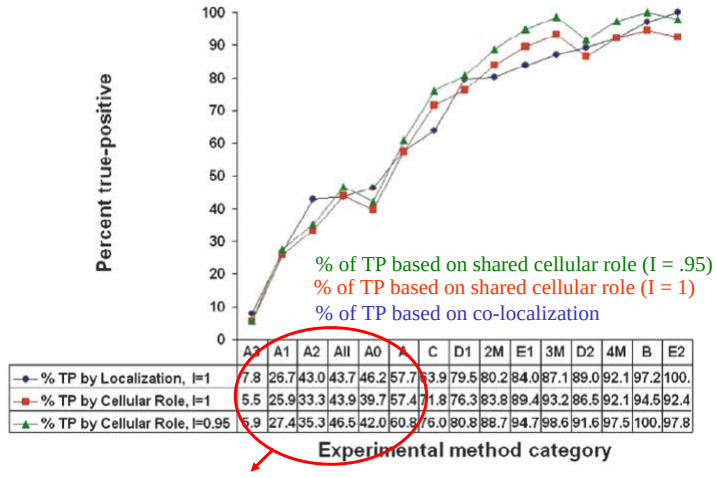
Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results)	516	53	33
A2: GY2H Ito *et al.* (core)	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

Large disagreement betw methods

- **GY2H: genome-scale Y2H**
- **2M, 3M, 4M: intersection of 2, 3, 4 methods**

Reliability of Protein Interaction D

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



TP = ~50%

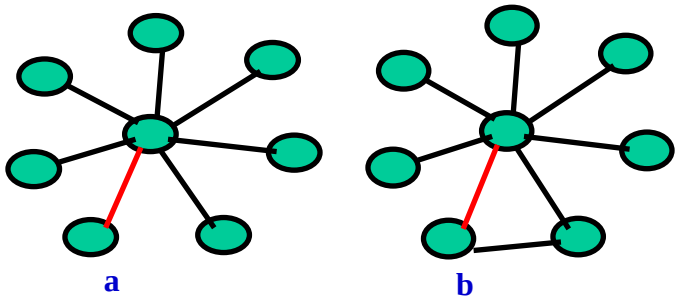
Objective

- **Some high-throughput protein interaction expts have as much as 50% false positives**
- **Can we find a way to rank candidate interaction pairs according to their reliability?**
- **How do we do this?**
  - Would knowing their neighbours help?
  - Would knowing their local topology help?
  - Would knowing their global topology help?

Would knowing their neighbours help?

The story of interaction generality

An Observation



- **It seems that configuration a is less likely than b in protein interaction networks**
- **Can we exploit this?**

Interaction Generality

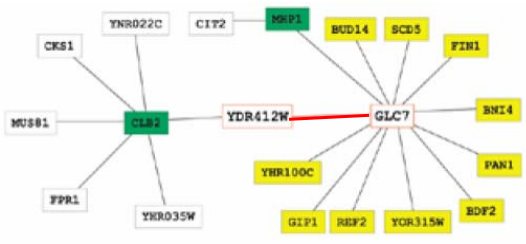
Saito et al., *NAR*, 30:1163-1168, 2002

Given an edge $X \leftrightarrow Y$ connecting two proteins, X and Y, the "interaction generality" measure $ig^{\mathcal{G}}(X \leftrightarrow Y)$ of this edge as defined as

$$ig^{\mathcal{G}}(X \leftrightarrow Y) = 1 + |\{X' \leftrightarrow Y' \in \mathcal{G} \mid X' \in \{X, Y\},\ deg^{\mathcal{G}}(Y') = 1\}|$$

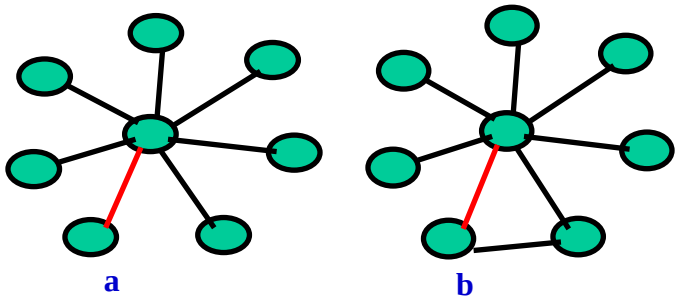
where $deg^{\mathcal{G}}(U) = |\{V|\ U \leftrightarrow V \in \mathcal{G}\}|$ is the degree of the node U in the undirected graph $\mathcal{G}$.

The number of proteins that "interact" with just X or Y, and nobody else



ig(YDR412W↔GLC7)
= 1 + # of yellow nodes

Assessing Reliability Using Interaction Generality



- **Recall configuration a is less likely than b in protein interaction networks**
- **The smaller the "ig" value of a candidate interaction pair is, the more likely that interaction is**

Evaluation wrt Intersection of Ito et al. & Uetz et al.

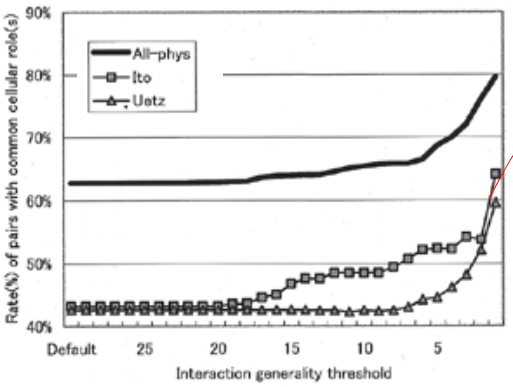
I.G.	Ito ol.	ovlap			Uetz ol.	ovlap		
1	229	66	34%	50%	236	58	29%	44%
2	137	34	54%	75%	226	37	57%	71%
3	57	16	63%	87%	113	16	71%	83%
4	43	6	69%	92%	66	6	79%	88%
5	24	4	73%	95%	38	5	83%	92%
6	16	1	75%	95%	37	2	88%	93%
7	27	0	79%	95%	20	3	90%	95%
8	23	1	83%	96%	16	2	92%	97%
9	9	1	84%	97%	4	0	93%	97%
10	2	0	84%	97%	44	0	98%	97%
11	0	0	84%	97%	9	2	99%	98%
12	1	0	84%	97%	4	0	100%	98%
13	13	0	86%	97%	0	1	100%	99%
14	15	0	89%	97%	1	1	100%	100%
15	16	0	91%	97%	0	0	100%	100%
16	30	3	95%	99%	1	0	100%	100%
17	6	1	96%	100%	0	0	100%	100%
18	20	0	99%	100%	0	0	100%	100%
19	2	0	100%	100%	0	0	100%	100%
20	3	0	100%	100%	0	0	100%	100%
21	0	0	100%	100%	0	0	100%	100%
22	0	0	100%	100%	0	0	100%	100%
23	0	0	100%	100%	0	0	100%	100%
24	0	0	100%	100%	0	0	100%	100%
25	0	0	100%	100%	0	0	100%	100%
26–	0	0	100%	100%	0	0	100%	100%
Total	673	133			815	133		

There are 229 pairs in Ito having ig = 1. Of these, 66 (or 34%) are also reported by Uetz

- **Interacting pairs c'mon to Ito et al. & Uetz et al. are more reliable**
- **Also have smaller "ig"**

⇒ **"ig" seems to work**

Evaluation wrt Co-localization

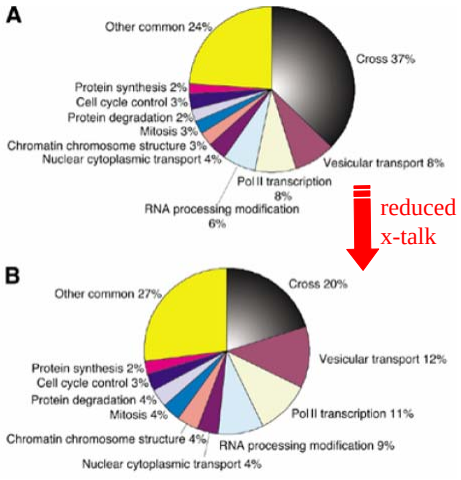


~60% of pairs in in Ito having ig=1 are known to have common localization

- **Interaction pairs having common cellular localization are more likely**
- **Also have lower "ig"**

⇒ **"ig" seems to work**

Evaluation wrt Co-cellular Role



- **Interaction pairs having common cellular role are more likely**
- **Also have lower "ig"**

⇒ **"ig" seems to work**

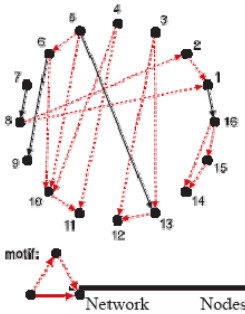
A: before restrict to pairs with "ig = 1"
B: after restrict to pairs with "ig = 1"

Would knowing their local topology help?

The story of interaction generality 2

Existence of Network Motifs

Milo et al., Science, 298:824-827, 2002



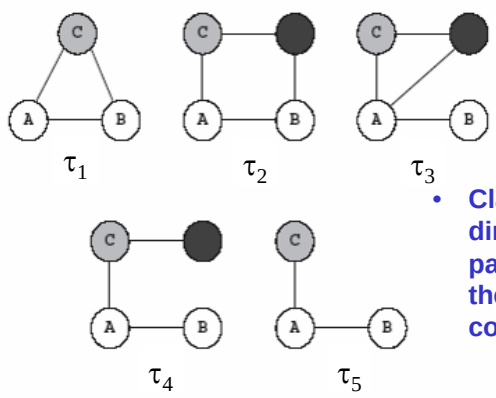
- **A network motif is just a local topological configuration of the network**
- **"Detected" in gene regulation networks, WWW links, etc.**

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)			X → Y → Z		Feed-forward loop	X, Y → Z, W		Bi-fan
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41

Observed 70 times in S. cerevisiae

Observed ~11 times in random data

5 Possible Network Motifs



- **Classify a protein C that directly interacts with the pair A↔B according to these 5 topological configurations**

A New Interaction Generality

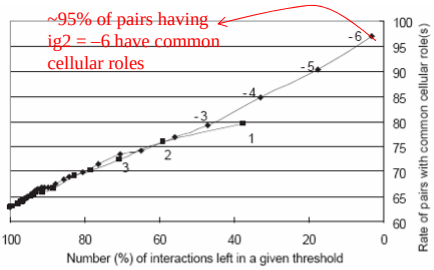
Saito et al., *Bioinformatics*, 19:756--763, 2003

The improved interaction generality measure $ig_2^{\mathcal{G}}(X \leftrightarrow Y)$ is defined as a weighted sum of the 5 local topological configurations $\tau_1, ..., \tau_5$ as

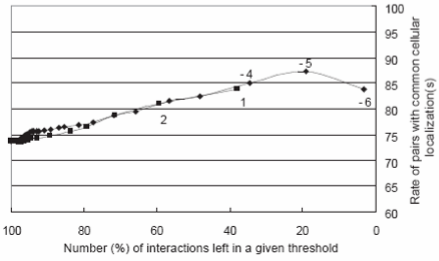
$$ig_2^{\mathcal{G}}(X \leftrightarrow Y) = \sum_{i=1}^{5} \lambda_i * |\{X' \mid X' \leftrightarrow Y' \in \mathcal{G},\ Y' \in \{X, Y\},\ \tau_i^{\mathcal{G}}(X', X \leftrightarrow Y)\}|$$

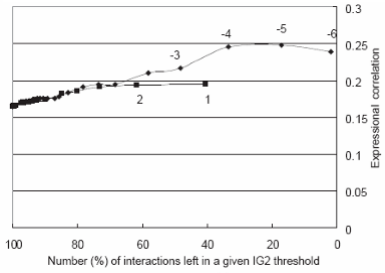
where λ_i is the weight for configuration τ_i, and $\tau_i^{\mathcal{G}}(X', X \leftrightarrow Y)$ means X' is in configuration τ_i in graph $\mathcal{G}$ wrt $X \leftrightarrow Y$.

Evaluation wrt Common Cellular Role, etc.



- **"ig_2" correlates well to common cellular roles, localization, & expression**
- **"ig_2" seems to work better than "ig"**





Would knowing their global topology help?

The story of interaction pathway reliability

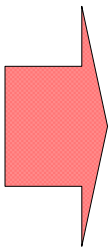
Some "Reasonable" Speculations

- **A true interacting pair is often connected by at least one alternative path (reason: a biological function is performed by a highly interconnected network of interactions)**
- **The shorter the alternative path, the more likely the interaction (reason: evolution of life is through "add-on" interactions of other or newer folds onto existing ones)**

Therefore...

Conjecture:

"An interaction that is associated with an alternate path of reliable interactions is likely to be reliable."



Idea:

Use **alternative interaction paths** as a measure to indicate functional linkage between the two proteins

Slide credit: See-Kiong Ng

Interaction Pathway Reliability

The "interaction pathway reliability" measure $ipr^{\mathcal{G}}(X \leftrightarrow Y)$ is defined as

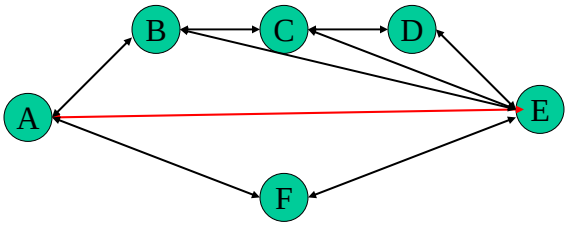
$$ipr^{\mathcal{G}}(X \leftrightarrow Y) = \max_{\phi \in \Phi^{\mathcal{G}}(X,Y)} \prod_{\{U \leftrightarrow V\} \in \phi} \left(1 - \frac{ig^{\mathcal{G}}(U \leftrightarrow V)}{ig^{\mathcal{G}}_{max}}\right)$$

where $ig^{\mathcal{G}}_{max} = \max\{ig^{\mathcal{G}}(X \leftrightarrow Y) \mid (X \leftrightarrow Y) \in \mathcal{G}\}$ is the maximum interaction generality value in $\mathcal{G}$; and $\Phi^{\mathcal{G}}(X,Y)$ is the set of all possible non-reducible paths between X and Y, but excluding the direct path $X \leftrightarrow Y$. Here, a path ϕ connecting X and Y is non-reducible if there is no shorter path ϕ' connecting X and Y that shares some common intermediate nodes with the path ϕ.

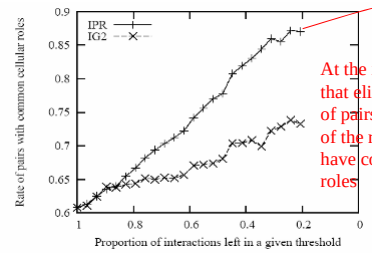
IPR is also called IRAP, "Interaction Reliability by Alternate Pathways"

Non-reducible Paths

- **Non-reducible paths are**
  - A←→F←→E
  - A←→B←→E
- **Reducible paths are**
  - A←→B←→C←→D←→E
  - A←→B←→C←→E



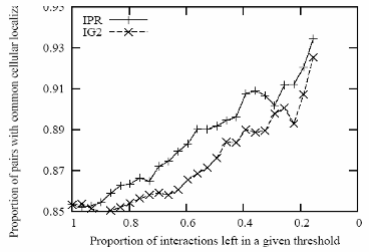
Evaluation wrt Common Cellular Role, etc

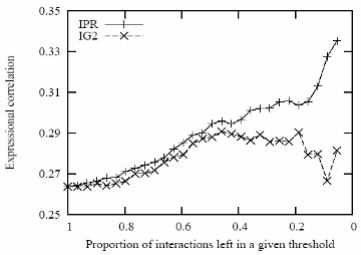


At the ipr threshold that eliminated 80% of pairs, ~85% of the of the remaining pairs have common cellular roles

- **"ipr" correlates well to common cellular roles, localization, & expression**

⇒ **"ipr" seems to work better than "ig_2"**

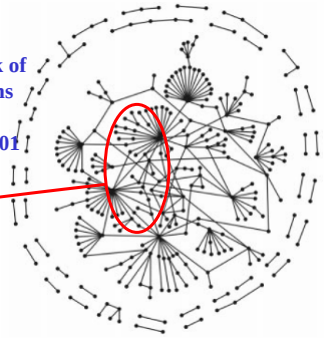


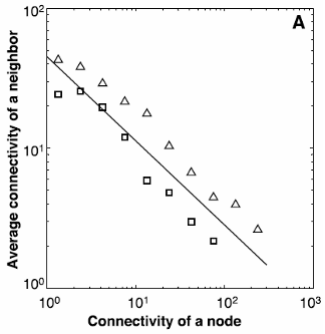


Stability in Protein Networks

Maslov & Sneppen, *Science*, 296:910-913, 2002

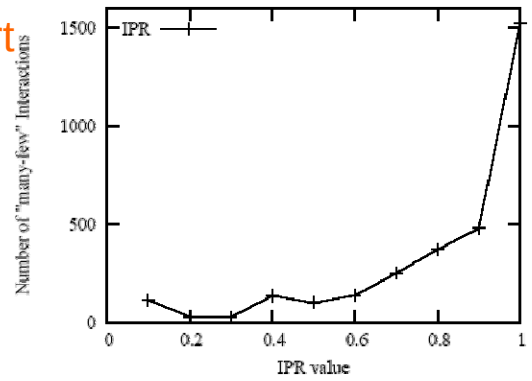
Part of the network of physical interactions reported by Ito et al., PNAS, 2001





- **According to Maslov & Sneppen**
  - Links betw high-connected proteins are suppressed
  - Links betw high- & low-connected proteins are favoured
- **This decreases cross talks & increases robustness**

Evaluation wrt "Many-few" Interactions



- **Number of "Many-few" interactions increases when more "reliable" IPR threshold is used to filter interactions**
- **Consistent with the Maslov-Sneppen prediction**

Evaluation wrt "Cross-Talkers"

- **A MIPS functional cat:**
  - | 02 | ENERGY
  - | 02.01 | glycolysis and gluconeogenesis
  - | 02.01.01 | glycolysis methylglyoxal bypass
  - | 02.01.03 | regulation of glycolysis & gluconeogenesis
- **First 2 digits is top cat**
- **Other digits add more granularity to the cat**

⇒ **Compare high- & low- IPR pairs that are not co-localised to determine number of pairs that fall into same cat. If more high-IPR pairs are in same cat, then IPR works**

Evaluation wrt "Cross-Talkers"

- **For top cat**
  - 148/257 high-IPR pairs are in same cat
  - 65/260 low-IPR pairs are in same cat
- **For fine-granularity cat**
  - 135/257 high-IPR pairs are in same cat.
    37/260 low-IPR pairs are in same cat

⇒ **IPR works**

⇒ **IPR pairs that are not co-localized are real cross-talkers!**

Example Cross Talkers

ProteinA	Cellular Localization	ProteinB	Cellular Localization	Functional Pathway
YDR299w	nucleolus-protein transport	YLR208w	cytoplasm-release of transport vesicles from ER	Vesicular transport (Golgi network)
YOL018c	endosome, ER-syntaxin SNARE	YMR117c	spindle pole body-spindle pole component	Cellular import
YDL154w	nucleus-recombination	YBR133c	cytoplasm- neg. regulator of kinase	Meiosis and budding
YGL192w	nucleus-put. Adenosine methyltransferase for sporulation	YBR057c	cytoplasm-meiosis potentially in premeiosis DNA synth	Development of asco-basido -zygo spore
YDR299w	nucleolous- protein transport	YPL085w	cytoplasm,ER-veiscle coat protein interacts cytoplasm, with sec23p	both in vesicular transport
YEL013w	vacuole-phosphorylated protein which interacts with Atg13p for cyto to vacuole targeting vacuole targeting	YFL039c	cytoskeleton-actin	Protein targeting and budding

Table 2
Examples of interactions with high IRAP values (≥ 0.95) between non-co-localized proteins ("cross-talkers") involved in the same cellular pathway

Can local topology do better?

The story of FS Weight

Guilt by Association of Common Interaction Partners

- **Two proteins that have a large proportion of their interaction partners in common are likely to directly interact also**
- **In fact, this is a special case of the "alternative paths" used in the IPR index, because length-1 alternative paths = shared interaction partners**

Czekanowski-Dice Distance

- **Functional distance between two proteins** (Brun et al, 2003)

$$D(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}$$

- N_k is the set of interacting partners of k
- X Δ Y is symmetric diff betw two sets X and Y
- Greater weight given to similarity

Is this a good measure if u and v have very diff number of neighbours?

⇒ **Similarity can be defined as**

$$S(u,v) = 1 - D(u,v) = \frac{2X}{2X + (Y + Z)}$$

Functional Similarity Estimate: FS-Weighted Measure

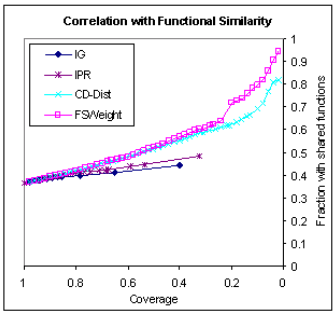
- **FS-weighted measure**

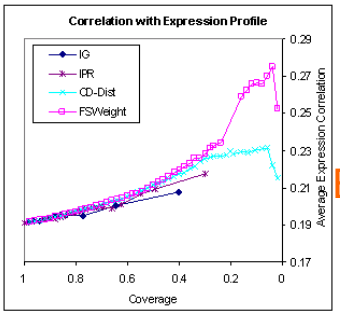
$$S(u,v) = \frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v|} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v|}$$

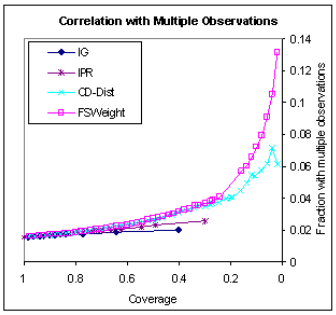
- N_k is the set of interacting partners of k
- Greater weight given to similarity

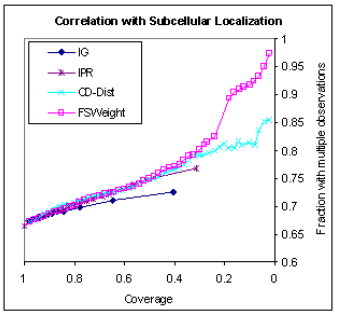
⇒ **Rewriting this as**

$$S(u,v) = \frac{2X}{2X + Y} \times \frac{2X}{2X + Z}$$







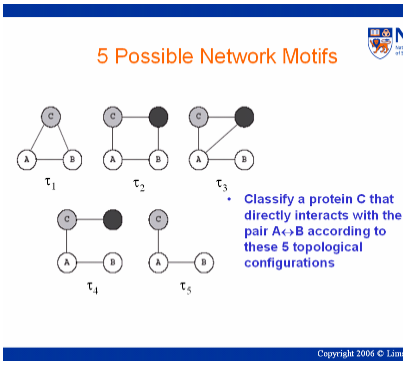


Evaluation wrt Common Cellular Role,etc

Another way to improve using local topology information

The story of meso-scale network motifs

Motivation for "Meso Scale"



- These motifs are very local and very small
- Many processes in biological network are ``meso-scale'' (5-25 proteins)

⇒ Maybe we should also use meso-scale motifs?

What is a network motif?

- A network motif g in a PPI network G is a connected unlabelled undirected topological pattern of inter-connections that is **repeated** and **unique** in G
- Repeated: f_g, the number of occurrences of g in G, is more than threshold F
- Unique: s_g, the number of times f_g exceeds $f_{g,rand,i}$ over total number of randomized networks considered, is more than threshold S

Example

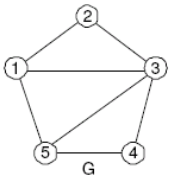


Figure 1: Example graph G.

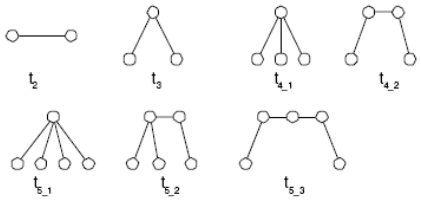


Figure 2: Size 2 to size 5 trees.

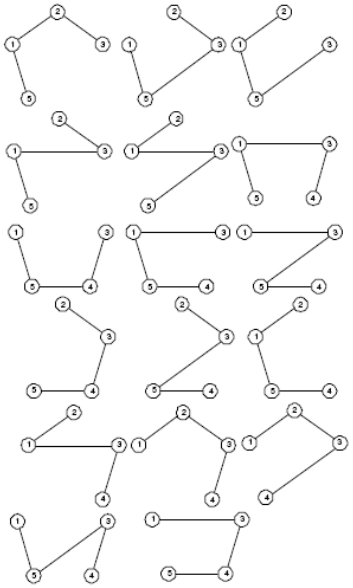
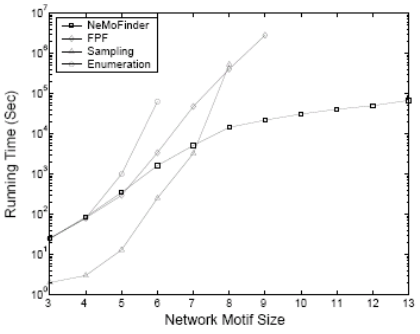
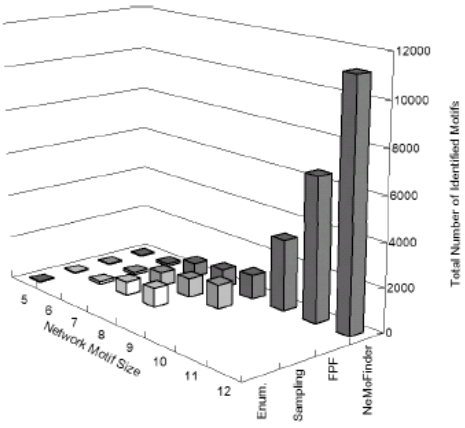


Figure 4: Occurrences of t_{4_2} in G.

NeMoFinder: Discovery of Meso-Scale Motifs





Motif Strength and PPI Reliability

- **Strength of a size k motif g is**

$$MS^k(g) = \frac{s_g \times f_g}{\max_k}$$

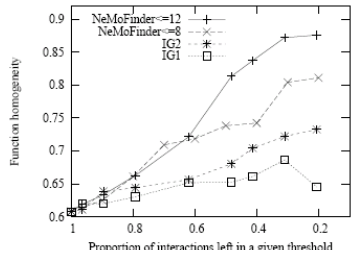
where $\max_k$ is max value of $s_g \times f_g$ over all size-k motifs

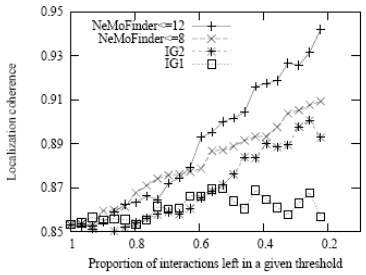
- **Motif-strength PPI reliability index is a pair of possibly interacting protein X ↔Y is**

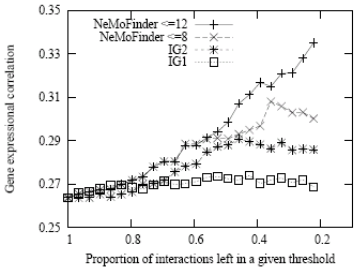
$$I(X \leftrightarrow Y) = \sum_{k=2}^{K} \sum_{i=0}^{n} MS^k(g_i) \times k$$

where g_i are motifs involving the edge X ↔Y, and k is size of g_i

Evaluation wrt Common Cellular Role, etc







- **Motif-strength PPI reliability index correlates well to common cellular roles, localization, & expression**

⇒ **works as well as "ipr"**

Some Observations

- **Meso-scale motifs are more reliable than small local motifs (c.f. "ig_2")**
- **Similar performance to "ipr", but may have advantages if network is sparse (i.e., where few alternate paths are present)**
- **Btw, this is the first time size-12 network motifs are known to be extracted from yeast PPI network**

Conclusions and Suggestions

Conclusions

- **There are latent local & global network "motifs" that indicate likelihood of PPIs**
- **These network "motifs" can be exploited in computational elimination of false positives from high-throughput PPI data**
- **FS-Weight, CD-Distance & meso-scale motifs are effective topologically-based computational measure for assessing the reliability (false positives) of PPIs**

Follow-Up Works

- **Expectation maximization can be applied on FS-Weight, CD-Distance, etc to further increase their power for detecting false positives**
- **FS-Weight, CD-Distance, etc can be adapted to detect false negatives**

Readings

- J. Chen et al, "Towards discovering reliable protein interactions from high-throughput experimental data using network topology", *Artificial Intelligence in Medicine*, 35:37-47, 2005
- J. Chen et al, "Increasing confidence of protein interactomes using network topological metrics", *Bioinformatics*, 22:1998-2004, 2006
- J. Chen et al, "NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs", *Proc. KDD 2006*
- H.N. Chua et al. "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions", *Bioinformatics*, 22:1623-1630, 2006
- H. N. Chua, L. Wong. "Increasing the Reliability of Protein Interactomes", *Drug Discovery Today*, 13(15/16):652-658, 2008
- G. Liu, J. Li, L. Wong. "Assessing and predicting protein interactions using both local and global network topological metrics", *Proc GIW2008*

Any Questions?