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 PPI data**
- **Identification of false positives**
  - – Interaction generality
  - – Interaction generality 2
  - – Interaction pathway reliability
  - – FS Weight
  - – Meso-scale network motifs
- **Identification of false negatives**
- **Uses of (cleansed) PPI data**
  - – Protein function prediction w/o homology info
  - – Protein complex prediction

How reliable are experimental protein-protein interaction data?

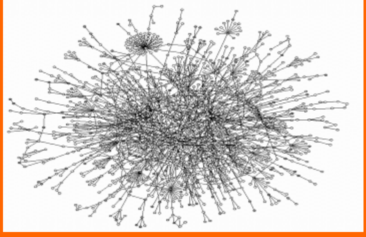


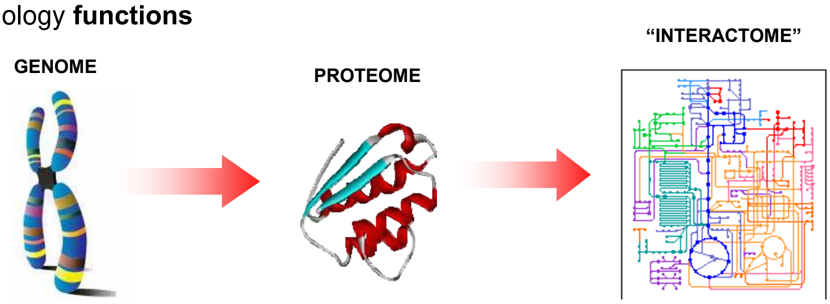
Figure credit: Jeong et al. 2001

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

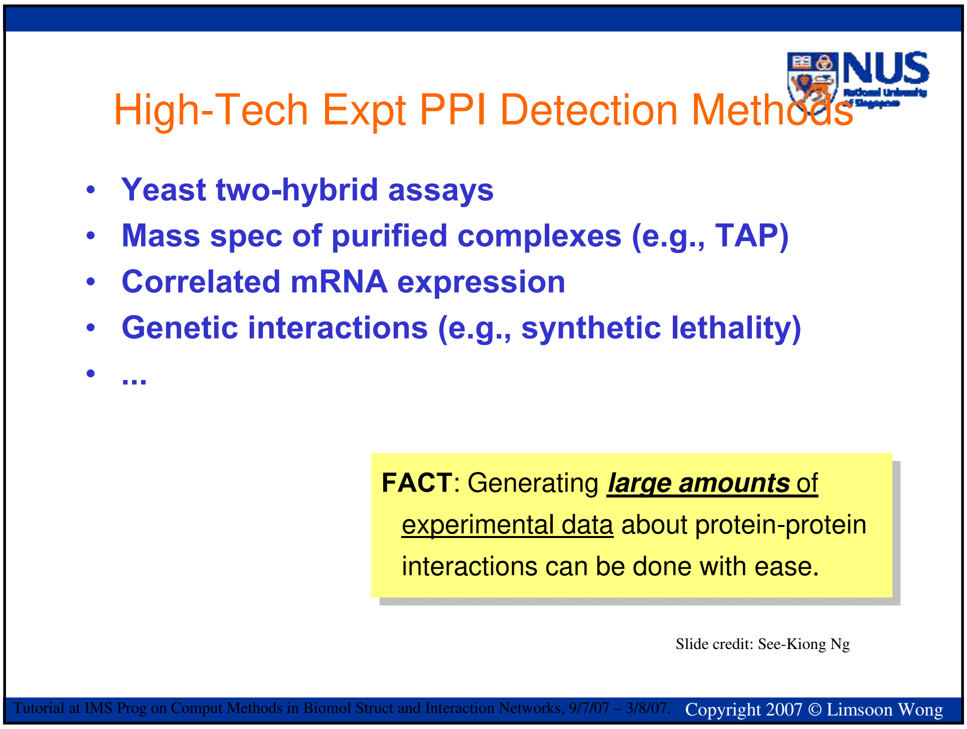


GENOME

PROTEOME

“INTERACTOME”

Slide credit: See-Kiong Ng

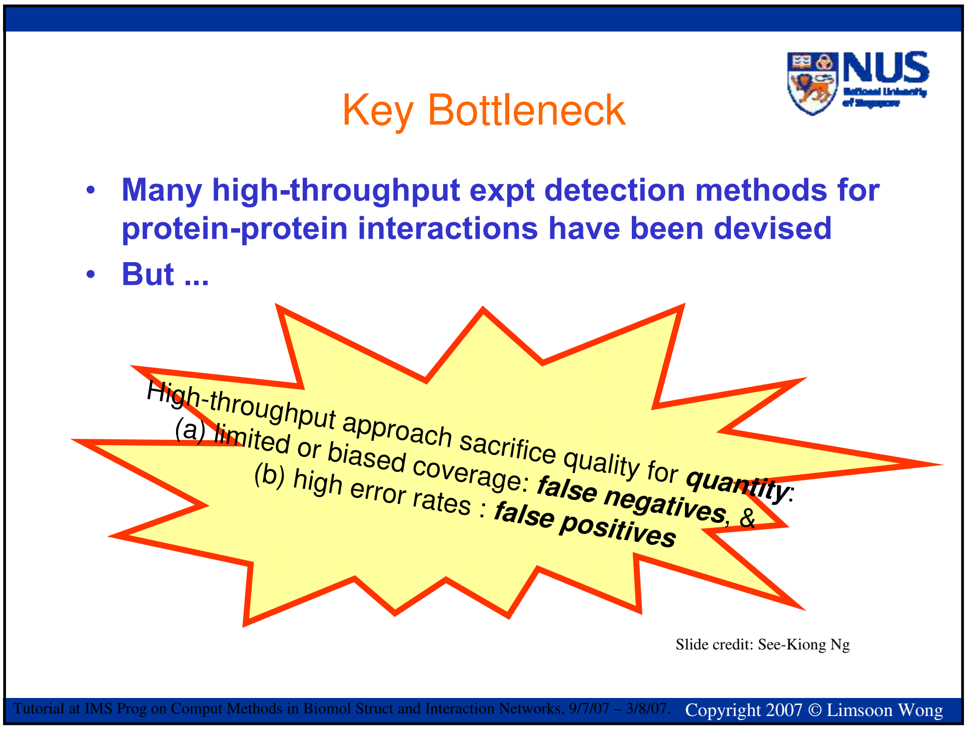


High-Tech Expt PPI Detection Methods

- 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

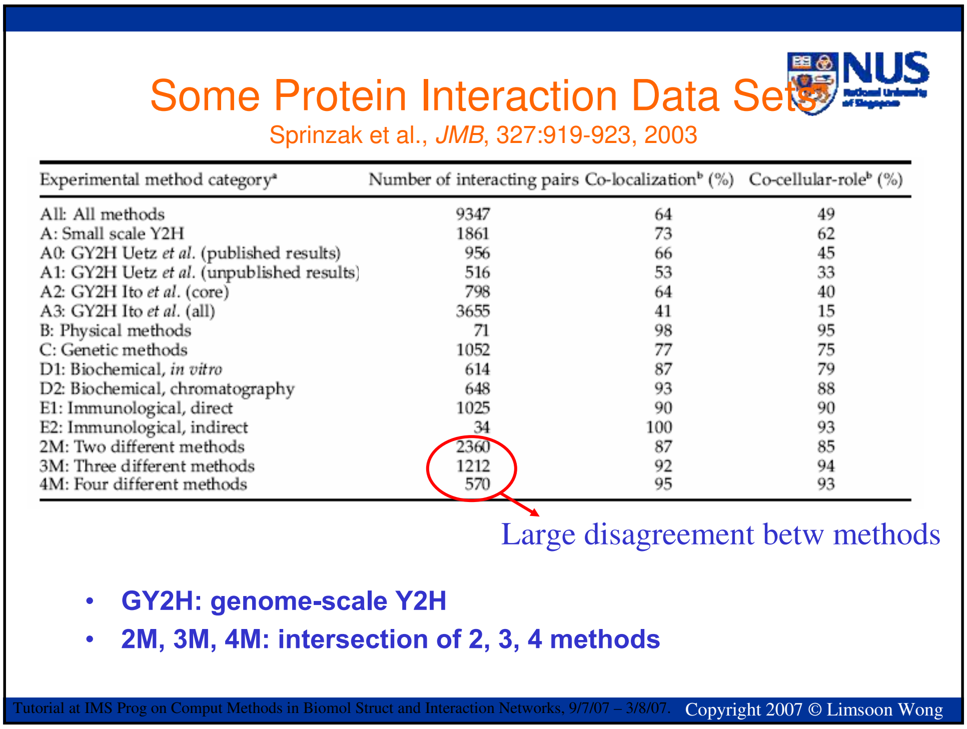


Key Bottleneck

- 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



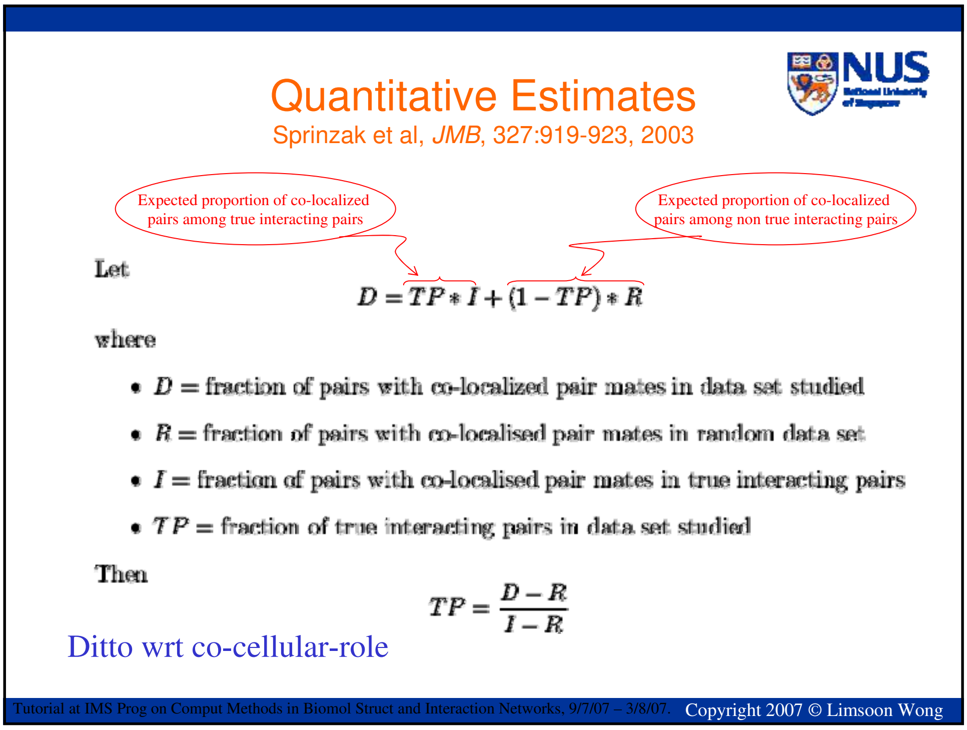
Some Protein Interaction Data Sets

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**



Quantitative Estimates

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

Let

$$D = \overbrace{TP * I}^{\text{Expected proportion of co-localized pairs among true interacting pairs}} + \overbrace{(1 - TP) * R}^{\text{Expected proportion of co-localized pairs among non true interacting pairs}}$$

where

- D = fraction of pairs with co-localized pair mates in data set studied
- R = fraction of pairs with co-localised pair mates in random data set
- I = fraction of pairs with co-localised pair mates in true interacting pairs
- TP = fraction of true interacting pairs in data set studied

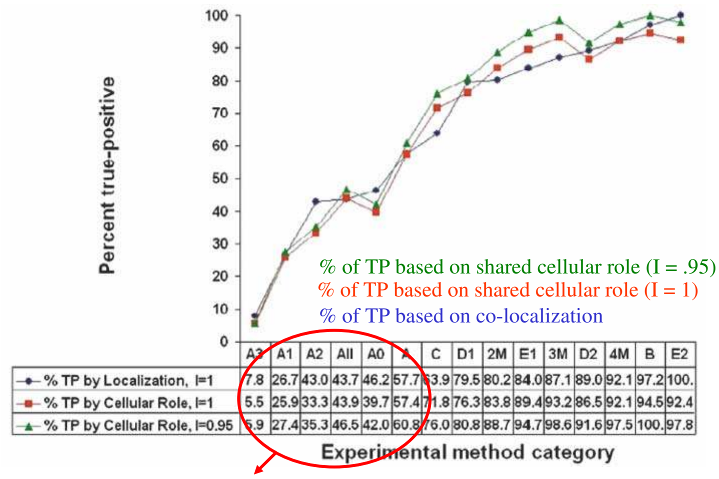
Then

$$TP = \frac{D - R}{I - R}$$

Ditto wrt co-cellular-role

Reliability of Protein Interaction Data

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



% of TP based on shared cellular role (I = .95)

% of TP based on shared cellular role (I = 1)

% of TP based on co-localization

	A3	A1	A2	AII	A0	A	C	D1	2M	E1	3M	D2	4M	B	E2
● % TP by Localization, I=1	7.8	26.7	43.0	43.7	46.2	57.7	63.9	79.5	80.2	84.0	87.1	89.0	92.1	97.2	100.
■ % TP by Cellular Role, I=1	5.5	25.9	33.3	43.9	39.7	57.4	71.8	76.3	83.8	89.4	93.2	86.5	92.1	94.5	92.4
▲ % TP by Cellular Role, I=0.95	5.9	27.4	35.3	46.5	42.0	60.8	76.0	80.8	88.7	94.7	98.6	91.6	97.5	100.	97.8

TP = ~50%

Are We There Yet?

	Coverage	Data quality
DNA genome sequence	**99% of genome sequence**	**99.9% correct**
mRNA profiling	**80-90% of transcripts represented**	**90% of spots are good data**
Protein interaction data	**10-30% of interactions catalogued**	**50-70% of interactions are spurious**

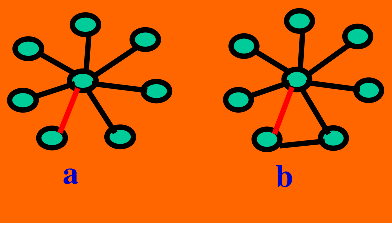
Slide credit: See-Kiong Ng

Objective

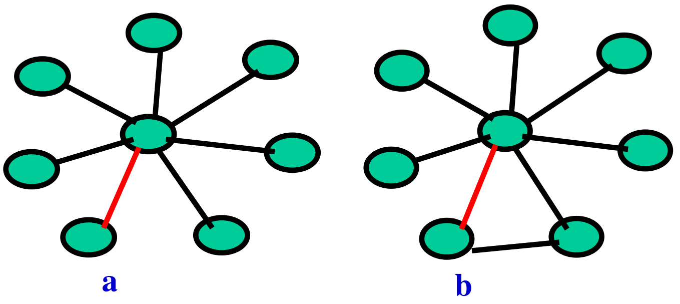
- **Some high-throughput protein interaction expts have as much as 50% false positives**
- **Can we find a way to rank candidate interaction pairs according to their reliability?**
- **How do we do this?**
  - Would knowing their neighbours help?
  - Would knowing their local topology help?
  - Would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

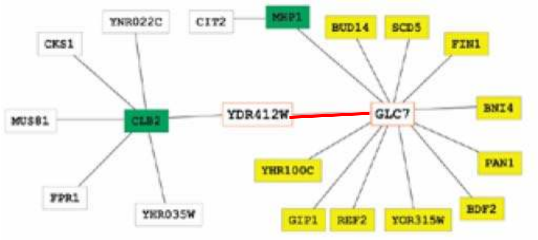
Saito et al., *NAR*, 30:1163-1168, 2002

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

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

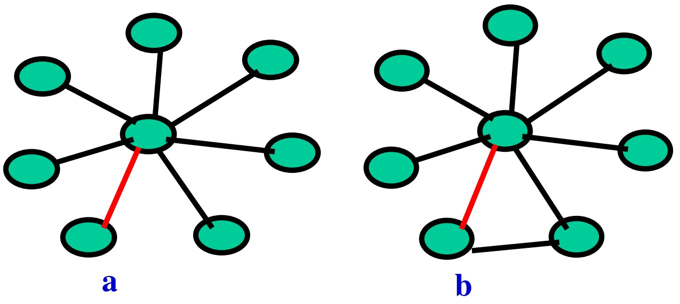
where $deg^{\mathcal{G}}(U) = |\{V| \ U \leftrightarrow V \ \in \mathcal{G}\}|$ is the degree of the node U in the undirected graph $\mathcal{G}$.

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



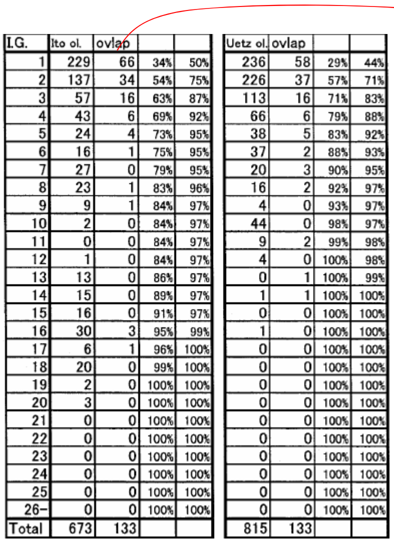
ig(YDR412W↔GLC7)
= 1 + # of yellow nodes

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.

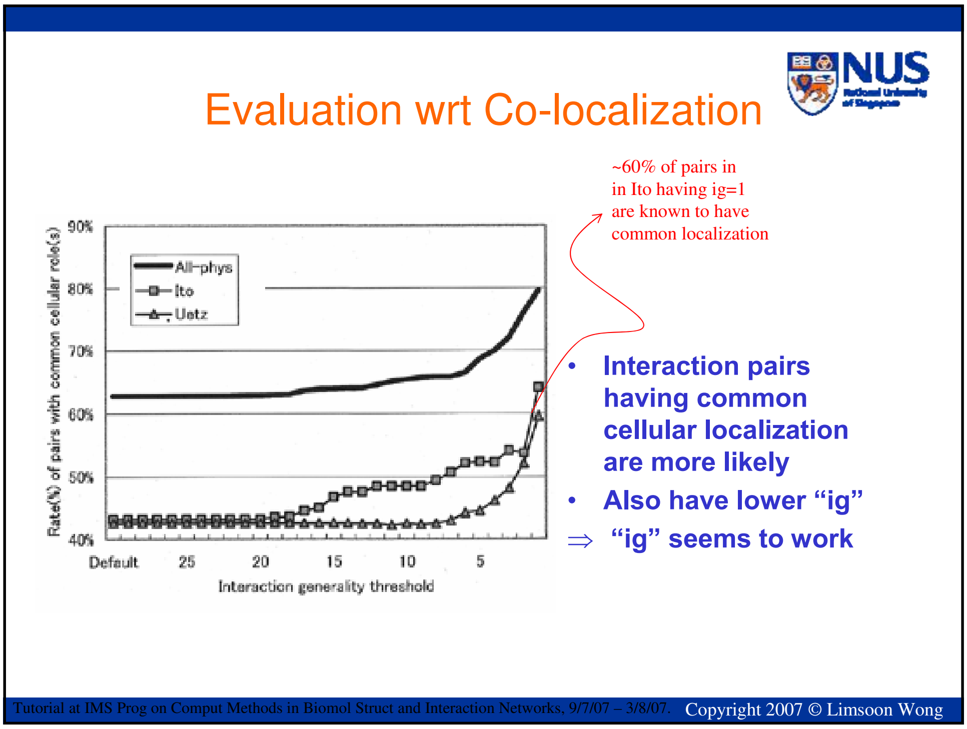


I.G.	Ito ol.	ovlap			Uetz ol.	ovlap		
1	229	66	34%	50%	236	58	29%	44%
2	137	34	54%	75%	226	37	57%	71%
3	57	16	63%	87%	113	16	71%	83%
4	43	6	69%	92%	66	6	79%	88%
5	24	4	73%	95%	38	5	83%	92%
6	16	1	75%	95%	37	2	88%	93%
7	27	0	79%	95%	20	3	90%	95%
8	23	1	83%	96%	16	2	92%	97%
9	9	1	84%	97%	4	0	93%	97%
10	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

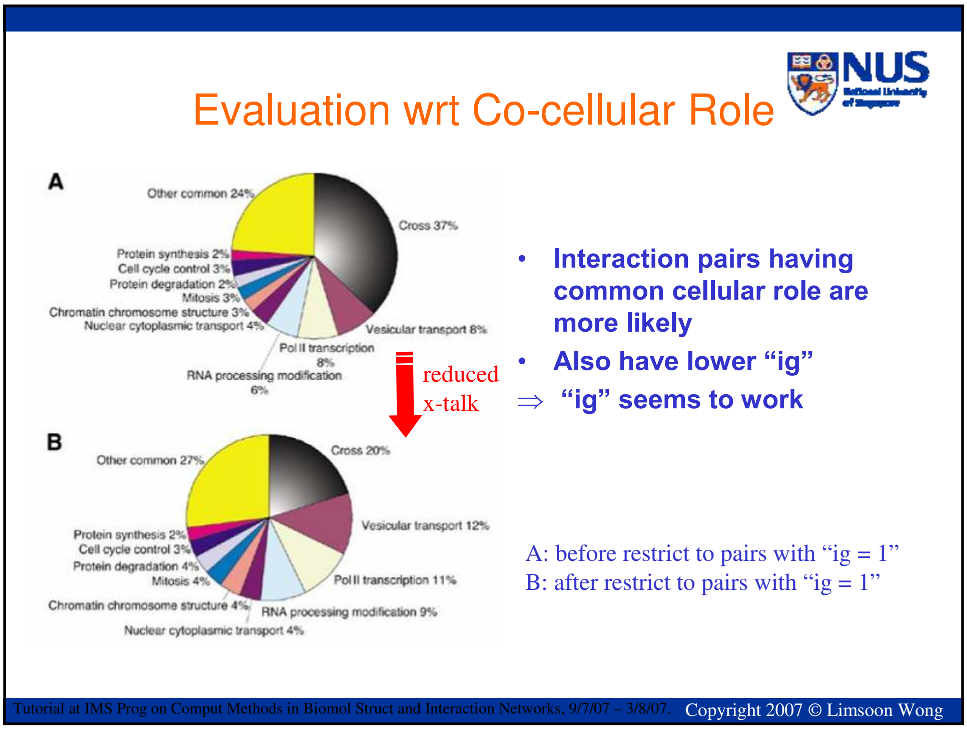


Evaluation wrt Co-localization

~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

reduced
x-talk

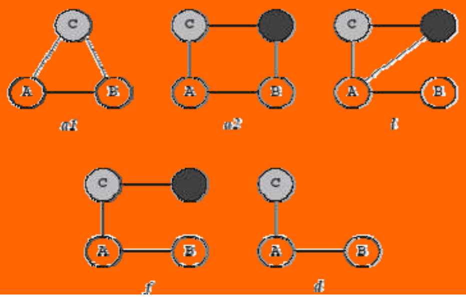
- Interaction pairs having common cellular role are more likely
- Also have lower “ig”

⇒ “ig” seems to work

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

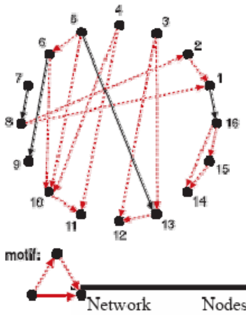
Would knowing their local topology help?

The story of interaction generality 2

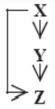
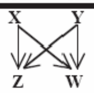


Existence of Network Motifs

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



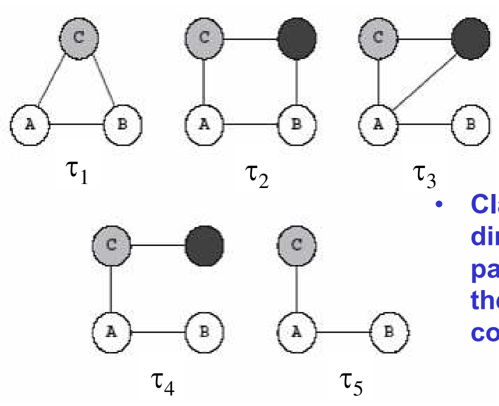
- **A network motif is just a local topological configuration of the network**
- **“Detected” in gene regulation networks, WWW links, etc.**

Network	Nodes	Edges	N_{real}	$N_{\text{rand}} \pm$ SD	Z score	N_{real}	$N_{\text{rand}} \pm$ SD	Z score
Gene regulation (transcription)					Feed-forward loop			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**

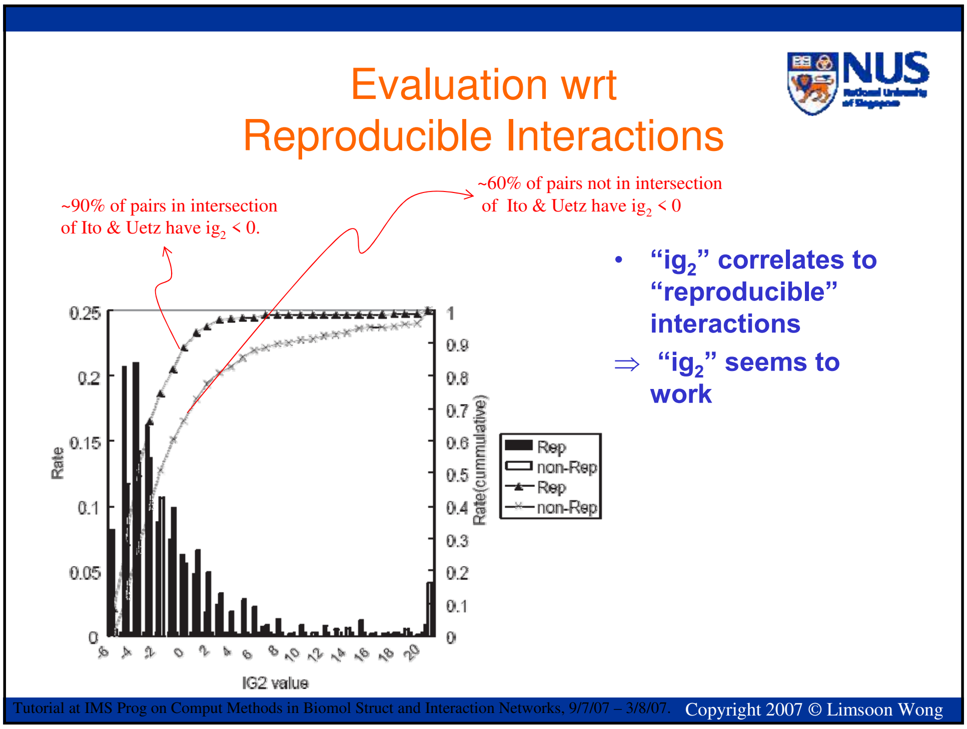
A New Interaction Generality

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 τ_1, ..., τ_5 as

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

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 $X \leftrightarrow Y$.



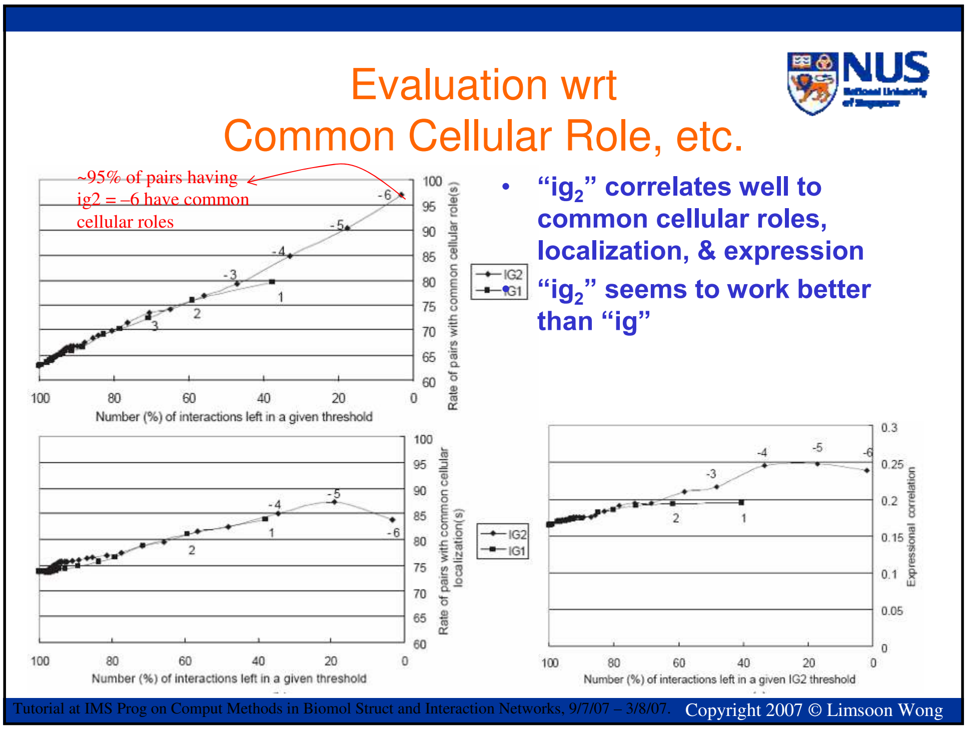
Evaluation wrt Reproducible Interactions

~90% of pairs in intersection of Ito & Uetz have ig_2 < 0.

~60% of pairs not in intersection of Ito & Uetz have ig_2 < 0

- “ig_2” correlates to “reproducible” interactions

⇒ “ig_2” seems to work



Evaluation wrt Common Cellular Role, etc.

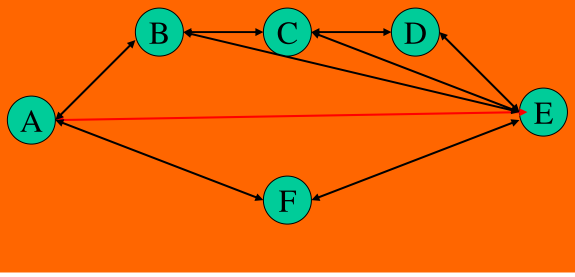
~95% of pairs having ig2 = –6 have common cellular roles

- “ig_2” correlates well to common cellular roles, localization, & expression

“ig_2” seems to work better than “ig”

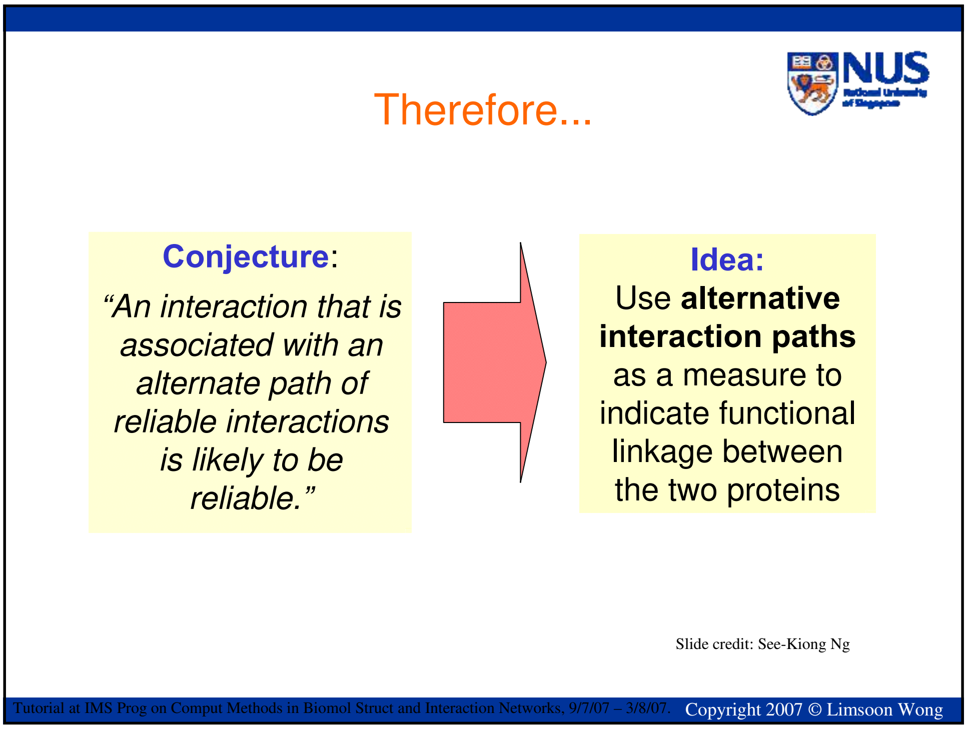
Would knowing their global topology help?

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...

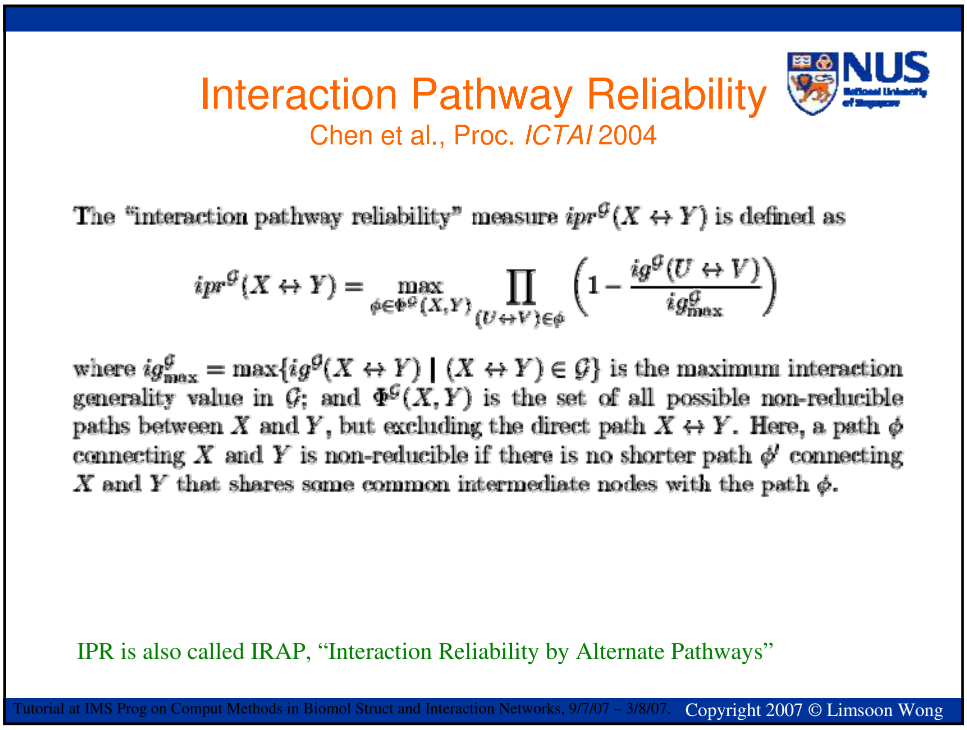
Conjecture:

"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

Chen et al., Proc. *ICTAI* 2004

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

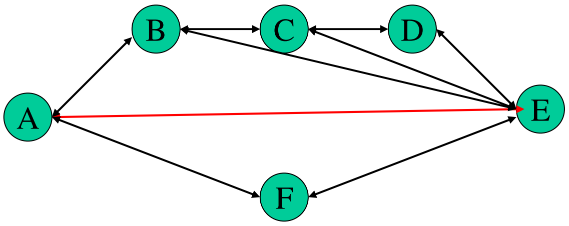
$$ipr^{\mathcal{G}}(X \leftrightarrow Y) = \max_{\phi \in \Phi^{\mathcal{G}}(X,Y)} \prod_{(U \leftrightarrow V) \in \phi} \left(1 - \frac{ig^{\mathcal{G}}(U \leftrightarrow V)}{ig^{\mathcal{G}}_{\max}}\right)$$

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

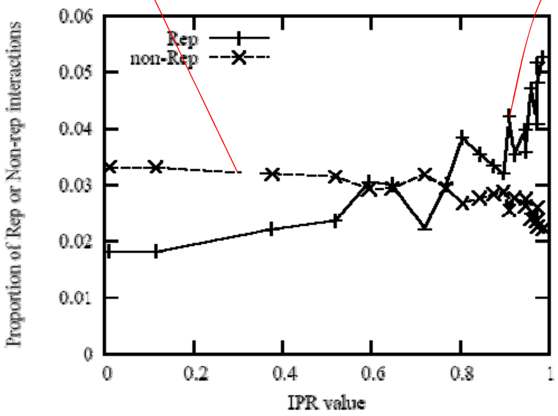
- **Non-reducible paths are**
  - A←→F←→E
  - A←→B←→E
- **Reducible paths are**
  - A←→B←→C←→D←→E
  - A←→B←→C←→E



Evaluation wrt Reproducible Interactions

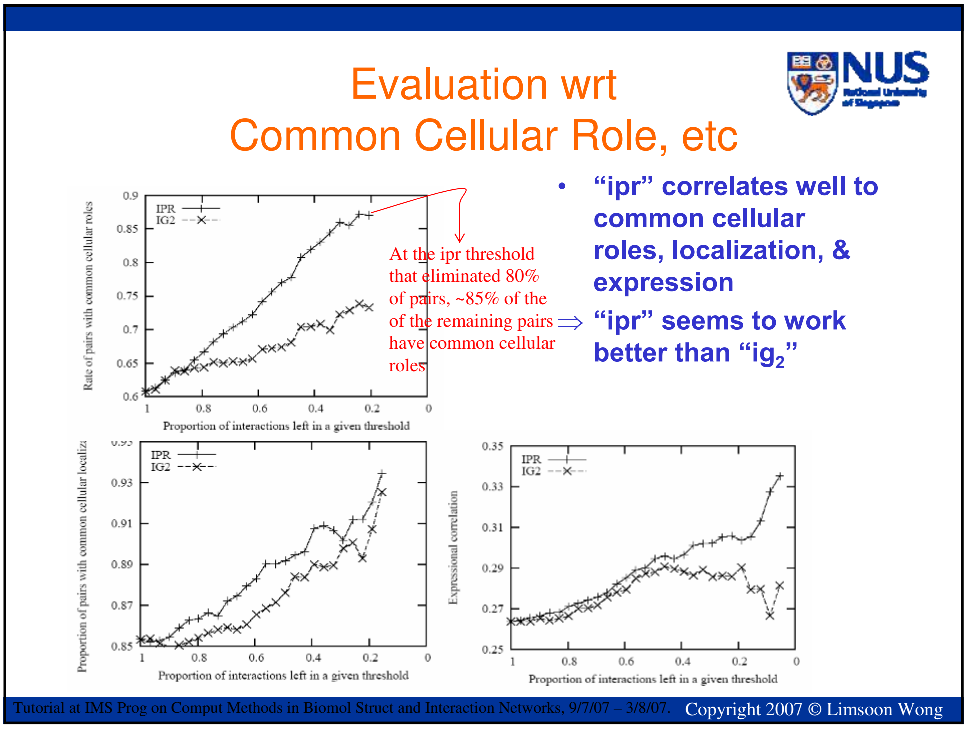
The number of pairs not in the intersection of Ito & Uetz is not changed much wrt the ipr value of the pairs

The number of pairs in the intersection of Ito & Uetz increases wrt the ipr value of the pairs



- **"ipr" correlates well to "reproducible" interactions**

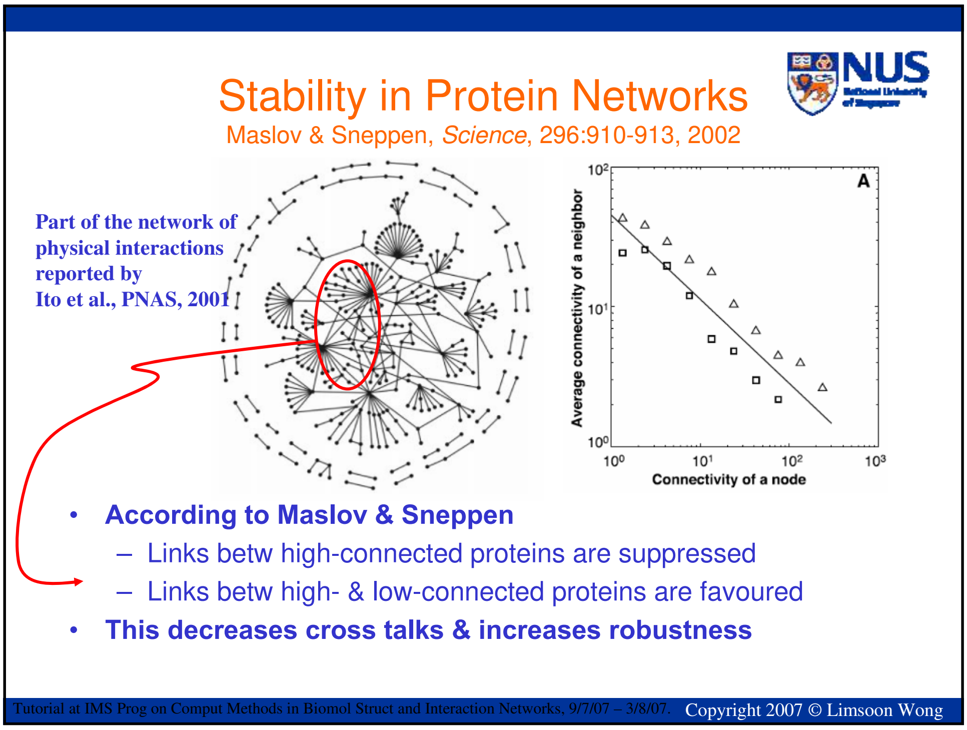
⇒ **"ipr" seems to work**



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$_2$”

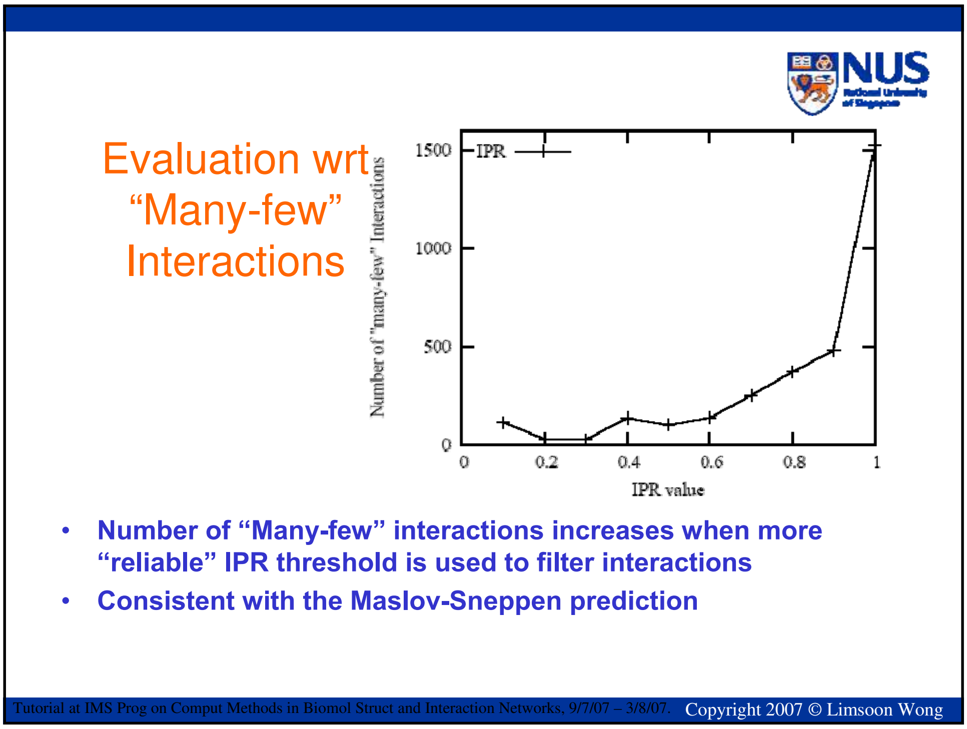


Stability in Protein Networks

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

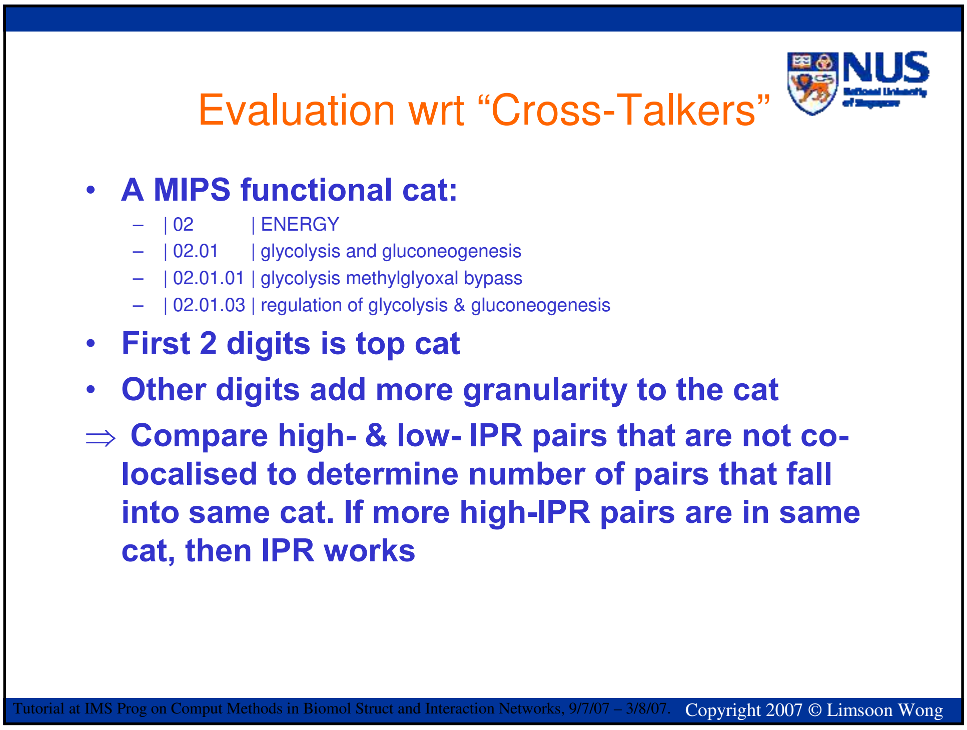
Part of the network of physical interactions reported by Ito et al., PNAS, 2001

- 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

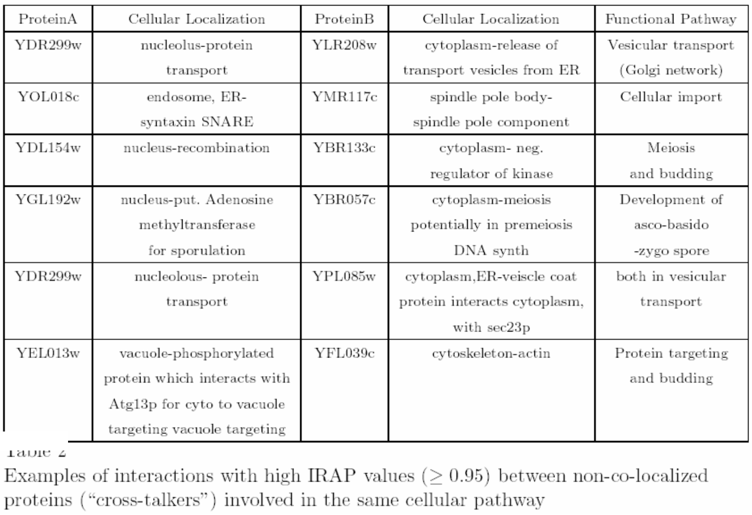
Evaluation wrt “Cross-Talkers”

- **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

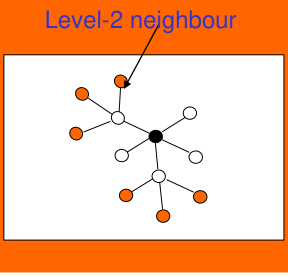


ProteinA	Cellular Localization	ProteinB	Cellular Localization	Functional Pathway
YDR299w	nucleolus-protein transport	YLR208w	cytoplasm-release of transport vesicles from ER	Vesicular transport (Golgi network)
YOL018c	endosome, ER-syntaxin SNARE	YMR117c	spindle pole body-spindle pole component	Cellular 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

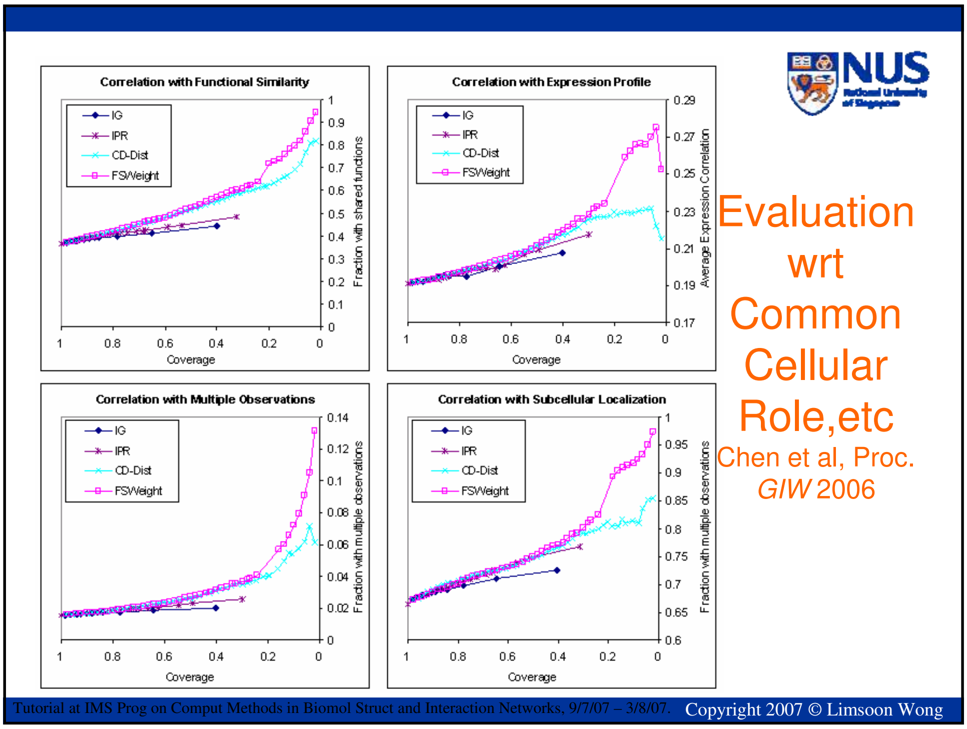
- **FS-weighted measure**

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

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

⇒ **Rewriting this as**

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



Correlation with Functional Similarity

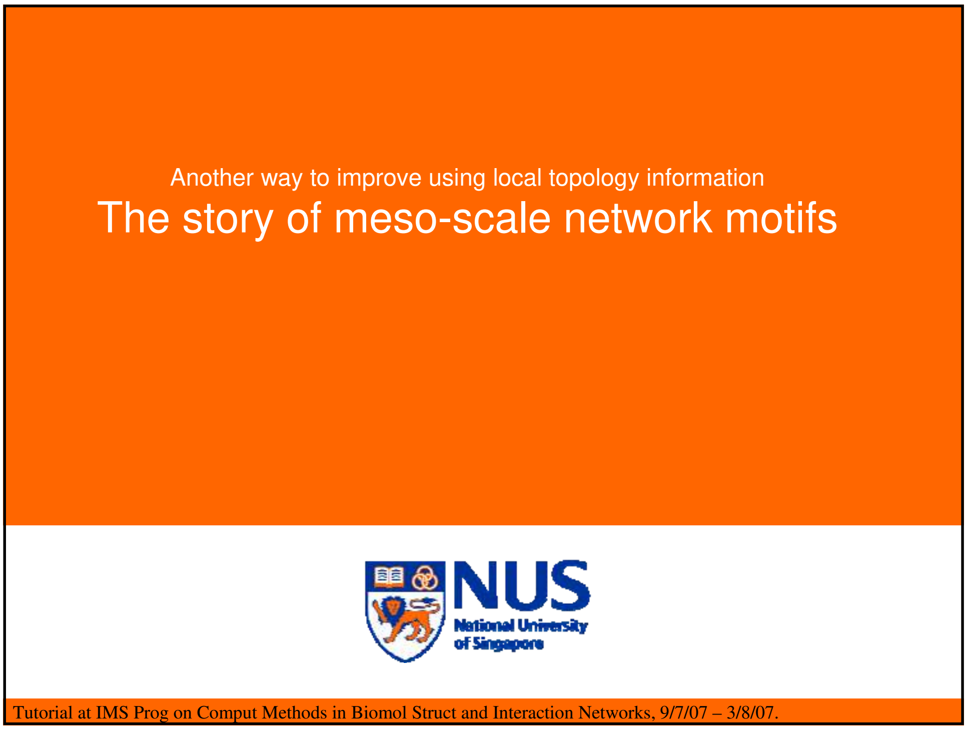
Correlation with Expression Profile

Correlation with Multiple Observations

Correlation with Subcellular Localization

Evaluation wrt Common Cellular Role,etc

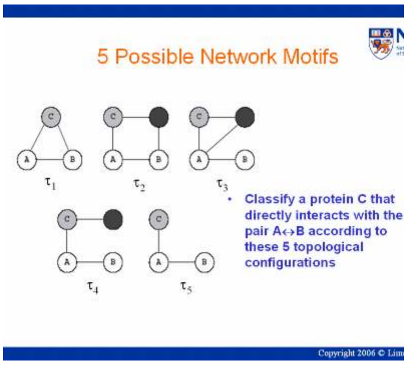
Chen et al, Proc. *GIW* 2006



Another way to improve using local topology information

The story of meso-scale network motifs

Motivation for "Meso Scale"



5 Possible Network Motifs

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

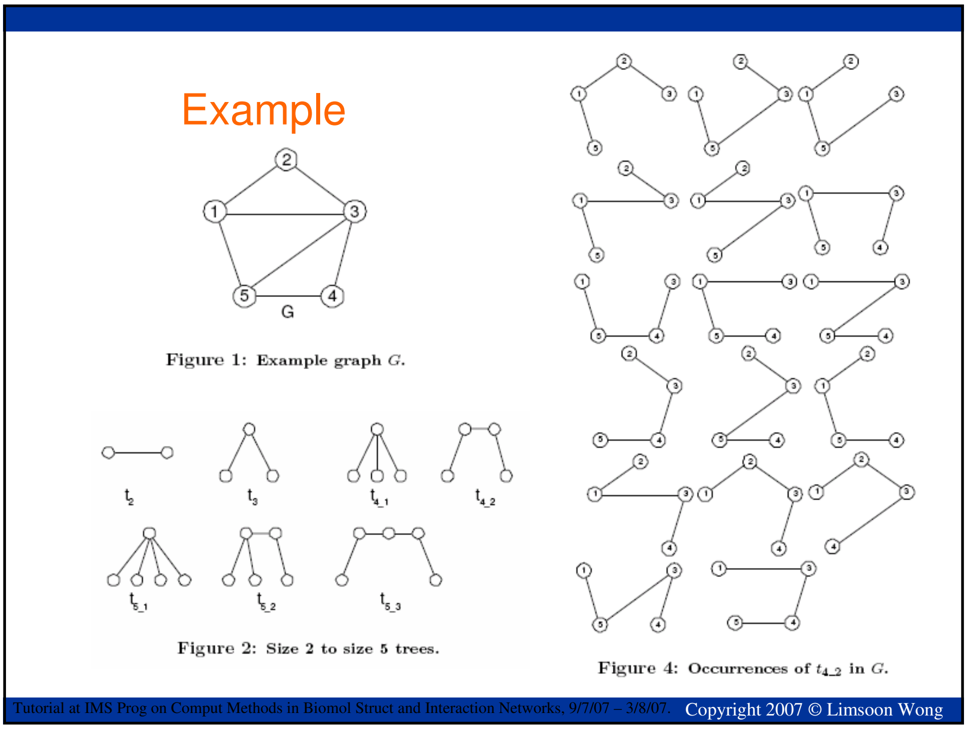
Copyright 2006 © Lim

- 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

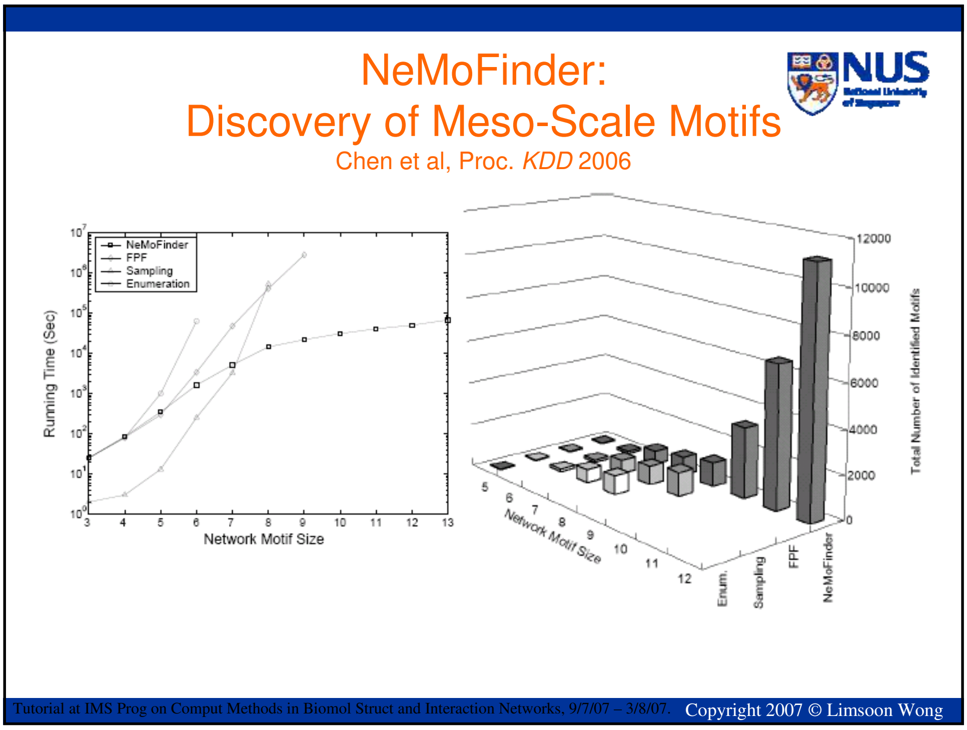


Example

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

Chen et al, Proc. *KDD* 2006

Motif Strength and PPI Reliability

- Strength of a size k motif g is

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

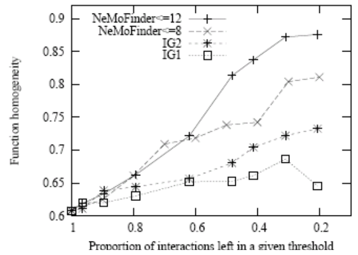
where $\max_k$ is max value of s_g x f_g over all size-k motifs

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

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

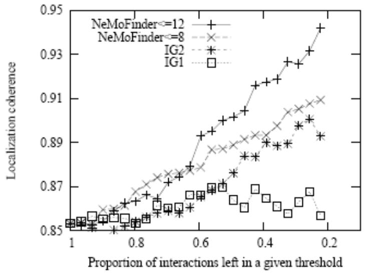
where g_i are motifs involving the edge X ↔Y, and k is size of g_i

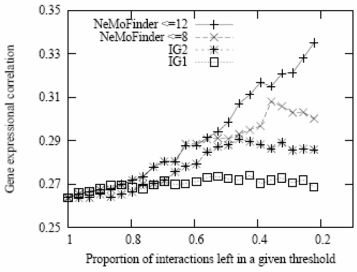
Evaluation wrt Common Cellular Role, etc

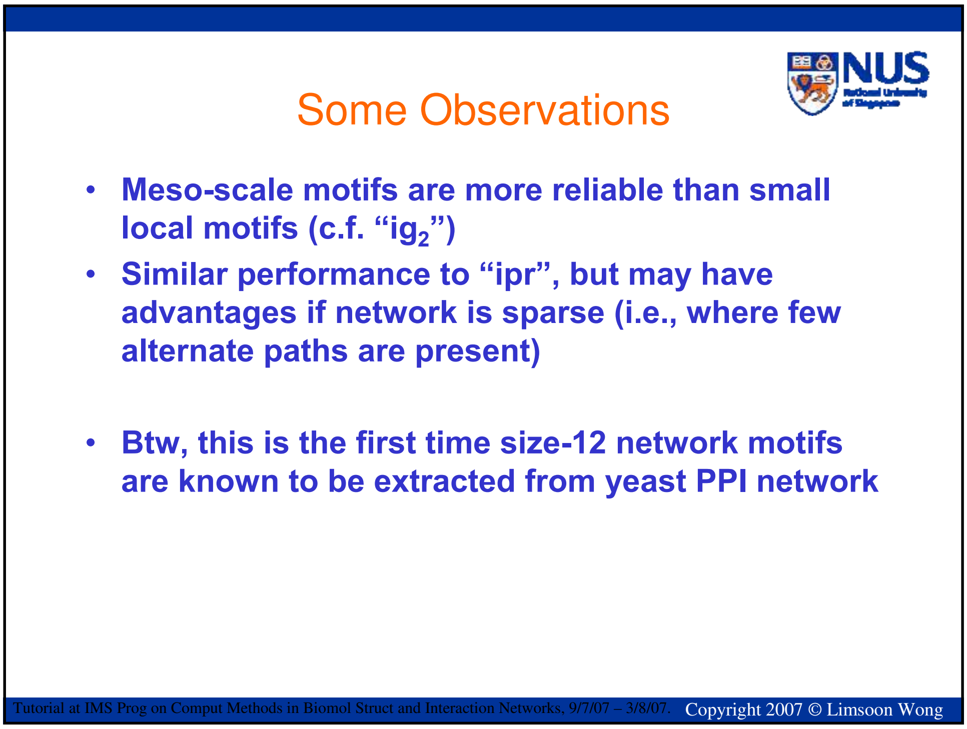


- Motif-strength PPI reliability index correlates well to common cellular roles, localization, & expression

⇒ works as well as "ipr"



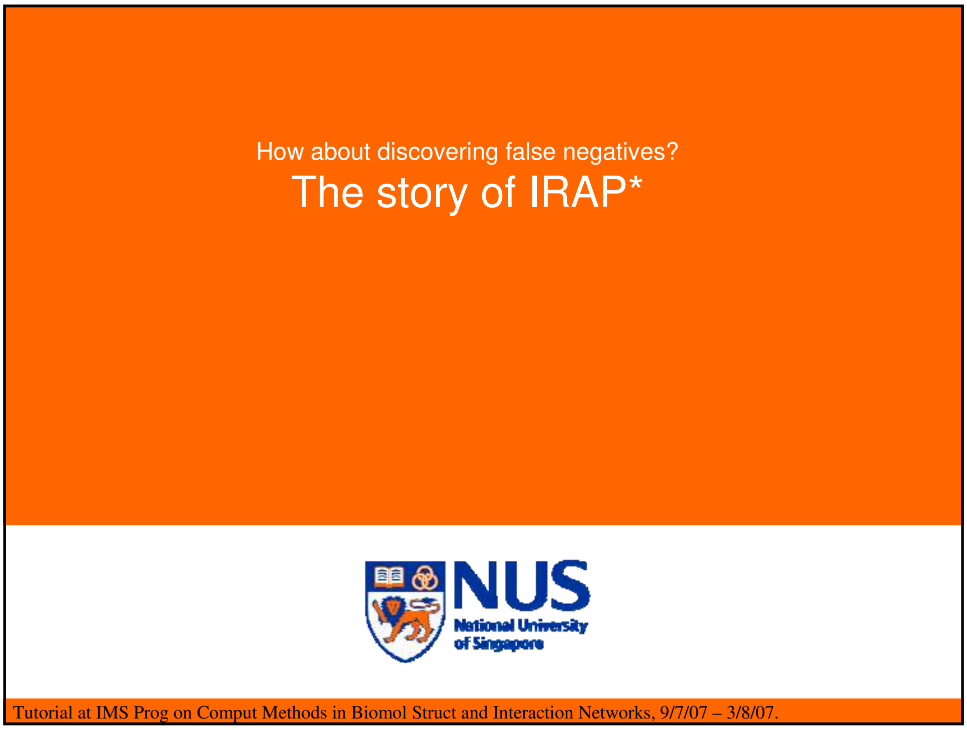




Some Observations

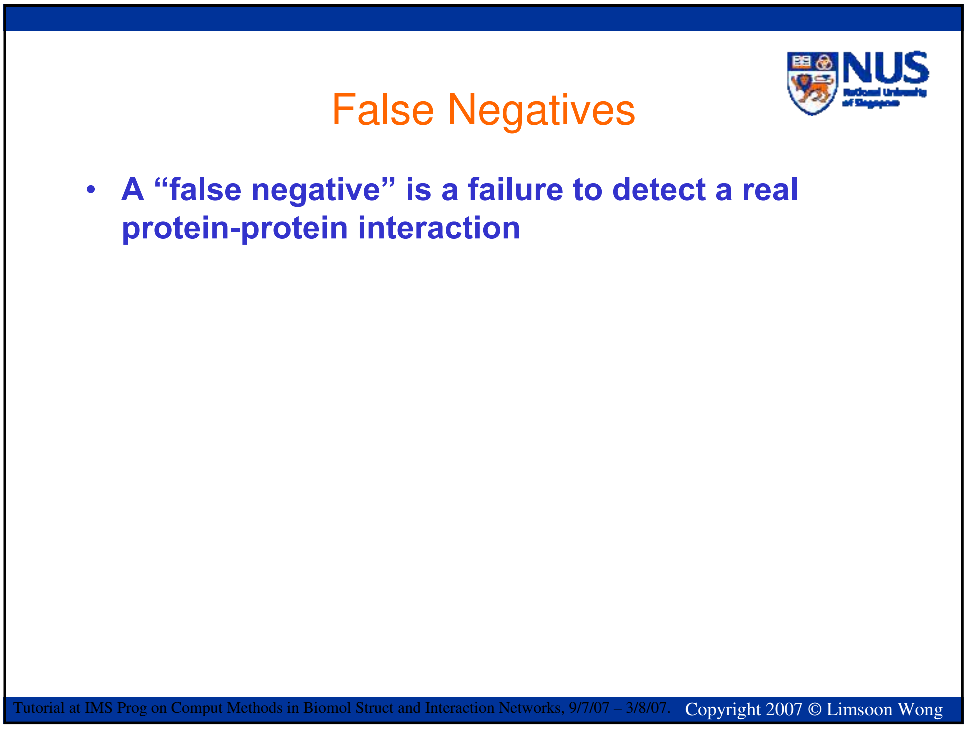
- **Meso-scale motifs are more reliable than small local motifs (c.f. “ig_2”)**
- **Similar performance to “ipr”, but may have advantages if network is sparse (i.e., where few alternate paths are present)**

- **Btw, this is the first time size-12 network motifs are known to be extracted from yeast PPI network**



How about discovering false negatives?

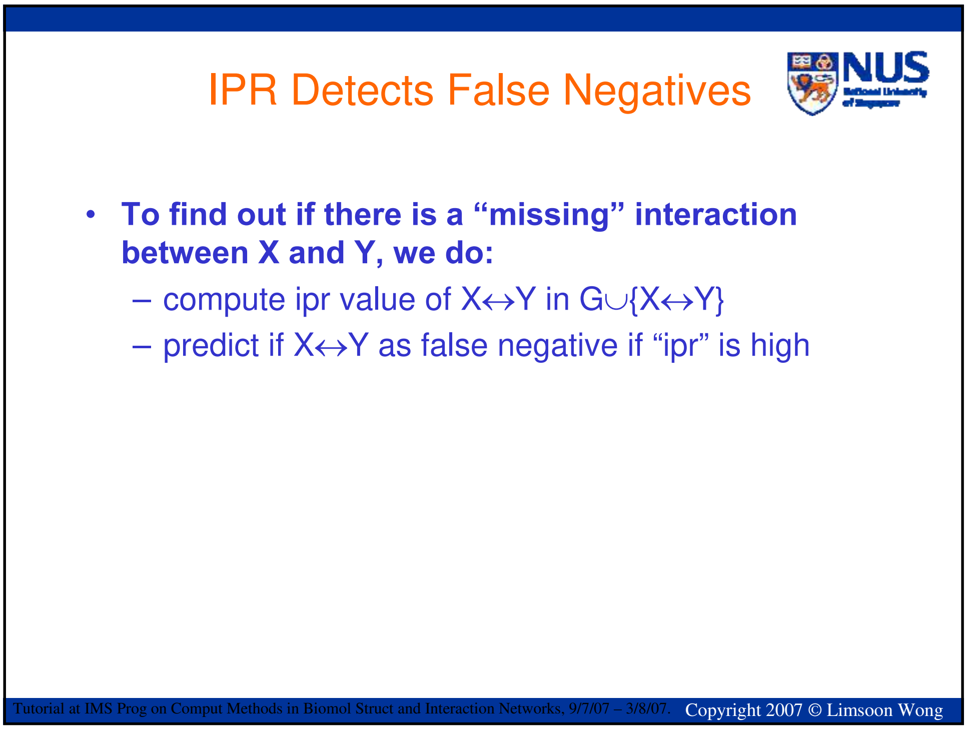
The story of IRAP*



False Negatives

- **A “false negative” is a failure to detect a real protein-protein interaction**

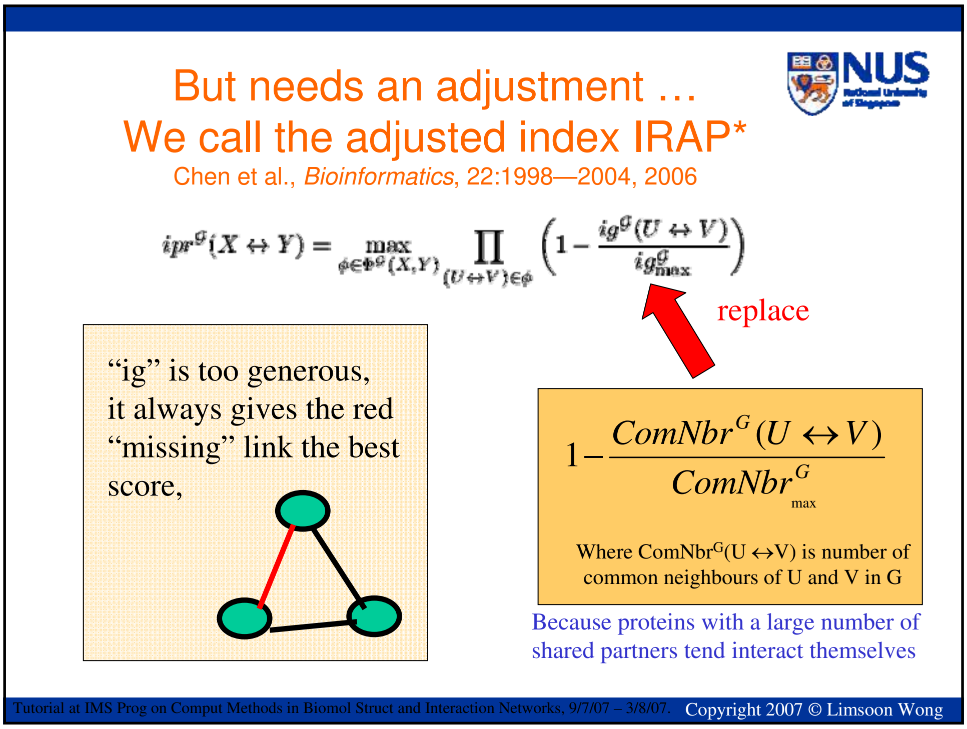
Tutorial at IMS Prog on Comput Methods in Biomol Struct and Interaction Networks, 9/7/07 – 3/8/07.



IPR Detects False Negatives

- **To find out if there is a “missing” interaction between X and Y, we do:**
  - compute ipr value of X↔Y in G∪{X↔Y}
  - predict if X↔Y as false negative if “ipr” is high

Tutorial at IMS Prog on Comput Methods in Biomol Struct and Interaction Networks, 9/7/07 – 3/8/07.



But needs an adjustment …
We call the adjusted index IRAP*

Chen et al., *Bioinformatics*, 22:1998—2004, 2006

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

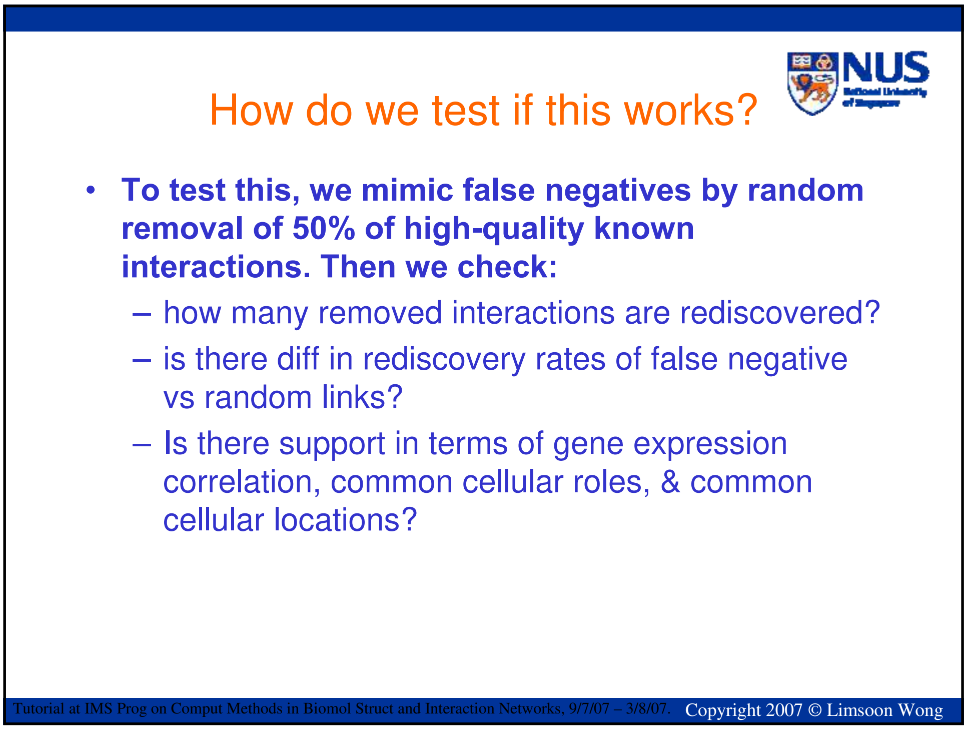
replace

"ig" is too generous, it always gives the red "missing" link the best score,

$$1 - \frac{ComNbr^{G}(U \leftrightarrow V)}{ComNbr^{G}_{\max}}$$

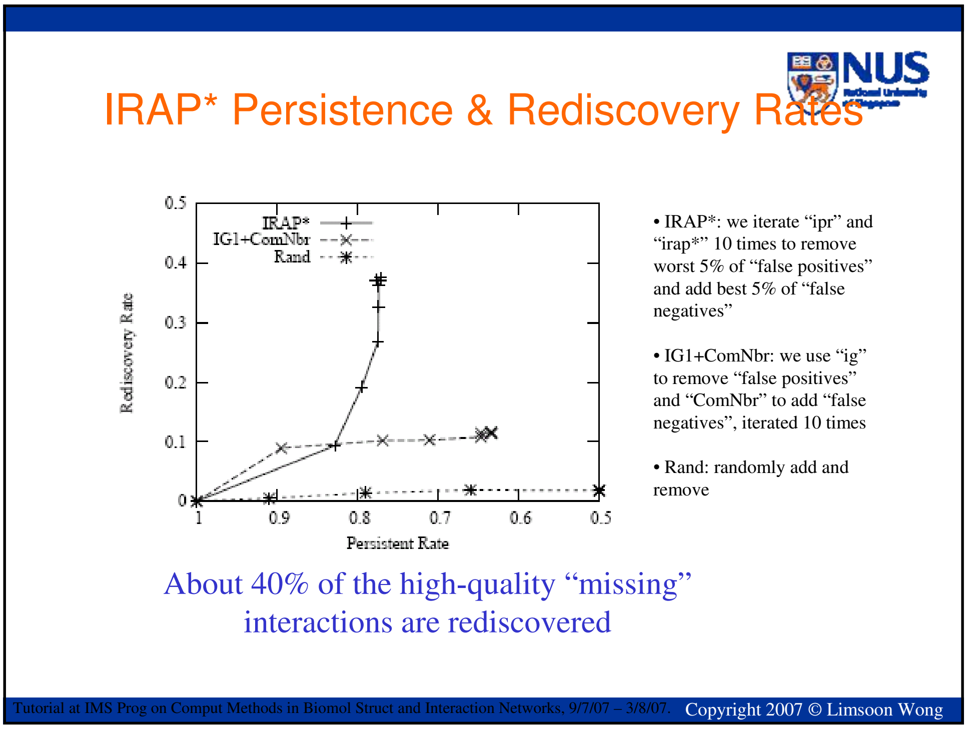
Where $ComNbr^{G}(U \leftrightarrow V)$ is number of common neighbours of U and V in G

Because proteins with a large number of shared partners tend interact themselves



How do we test if this works?

- **To test this, we mimic false negatives by random removal of 50% of high-quality known interactions. Then we check:**
  - how many removed interactions are rediscovered?
  - is there diff in rediscovery rates of false negative vs random links?
  - Is there support in terms of gene expression correlation, common cellular roles, & common cellular locations?

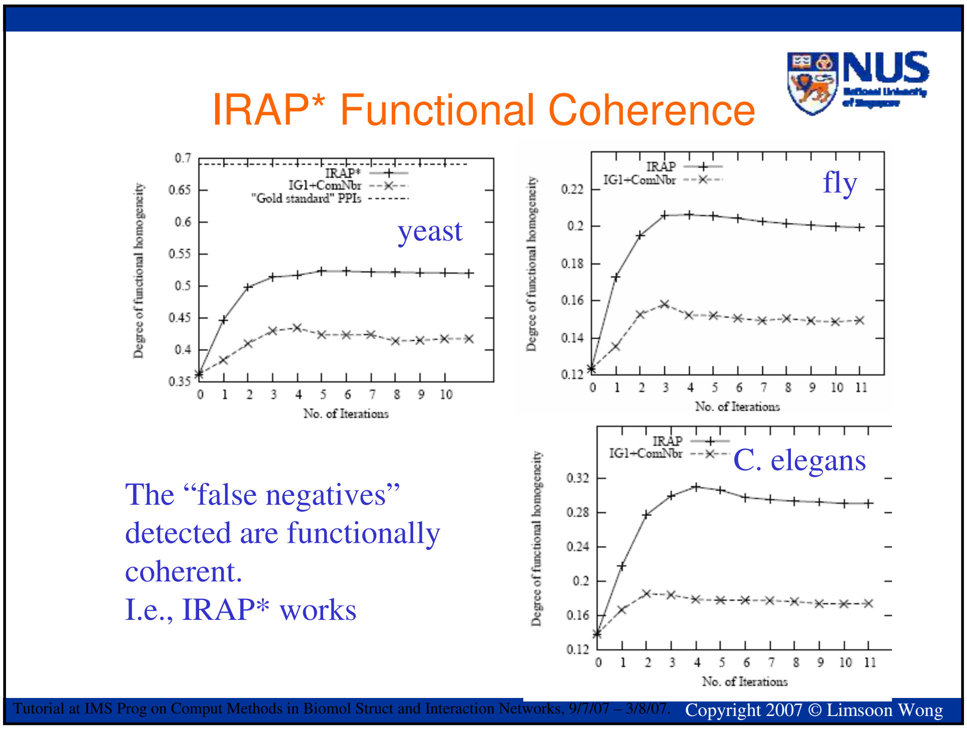


IRAP* Persistence & Rediscovery Rates

- IRAP*: we iterate “ipr” and “irap*” 10 times to remove worst 5% of “false positives” and add best 5% of “false negatives”
- IG1+ComNbr: we use “ig” to remove “false positives” and “ComNbr” to add “false negatives”, iterated 10 times
- Rand: randomly add and remove

About 40% of the high-quality “missing” interactions are rediscovered

Tutorial at IMS Prog on Comput Methods in Biomol Struct and Interaction Networks, 9/7/07 – 3/8/07. Copyright 2007 © Limsoon Wong



IRAP* Functional Coherence

The “false negatives” detected are functionally coherent.
I.e., IRAP* works

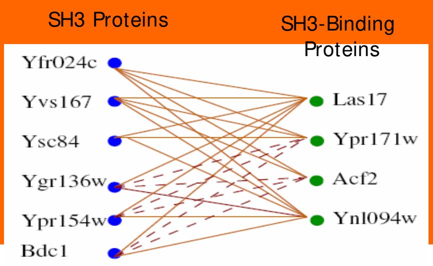
Tutorial at IMS Prog on Comput Methods in Biomol Struct and Interaction Networks, 9/7/07 – 3/8/07. Copyright 2007 © Limsoon Wong

Conclusions

- There are latent local & global network “motifs” that indicate likelihood of protein interactions
- These network “motifs” can be exploited in computational elimination of false positives & false negatives from high-throughput Y2H expt & possibly other highly erroneous interaction data
- IPR & meso-scale motifs are the most effective topologically-based computational measure for assessing the reliability (false positives) of protein-protein interactions detected by high-throughput methods
- IPR/IRAP* can discover new interactions (false negatives) not detected in the expt PPI network

Now that we have more reliable PPI networks, what can we do with them?

Protein function prediction w/o sequence homology information

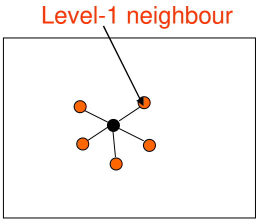


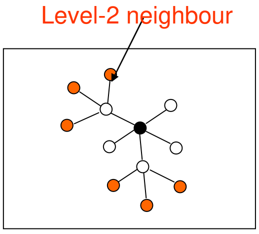
Protein Interaction Based Approaches

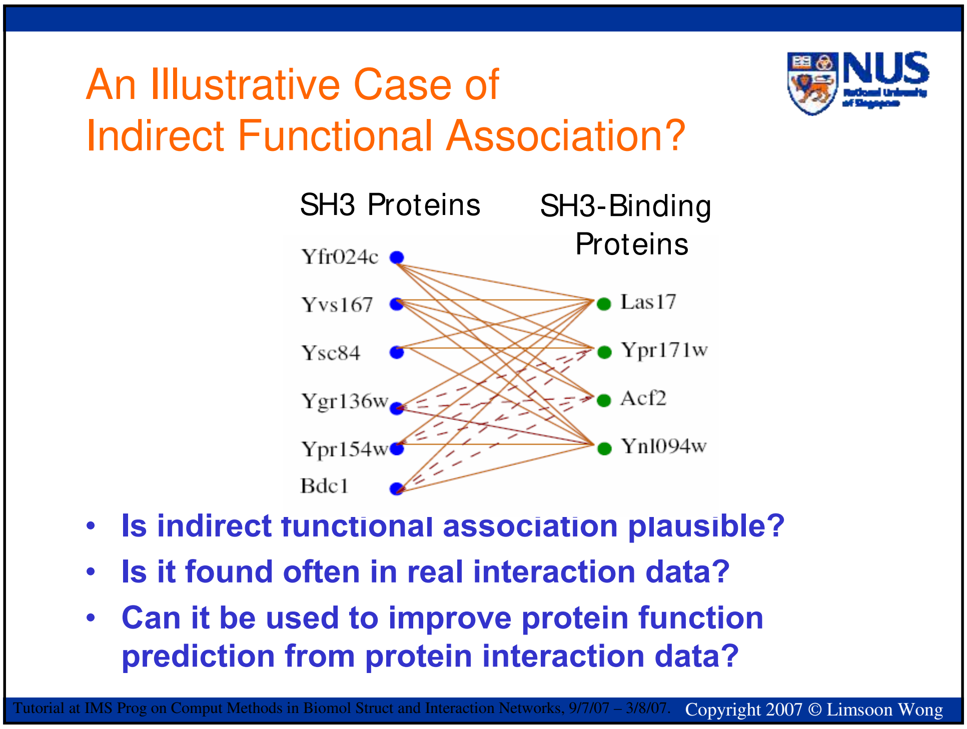
- **Neighbour counting** (Schwikowski et al, 2000)
  - Rank function based on freq in interaction partners
- **Chi-square** (Hishigaki et al, 2001)
  - Chi square statistics using expected freq of functions in interaction partners
- **Markov Random Fields** (Deng et al, 2003; Letovsky et al, 2003)
  - Belief propagation exploit unannotated proteins for prediction
- **Simulated Annealing** (Vazquez et al, 2003)
  - Global optimization by simulated annealing
  - Exploit unannotated proteins for prediction
- **Clustering** (Brun et al, 2003; Samanta et al, 2003)
  - Functional distance derived from shared interaction partners
  - Clusters based on functional distance represent proteins with similar functions
- **Functional Flow** (Nabieva et al, 2004)
  - Assign reliability to various expt sources
  - Function "flows" to neighbour based on reliability of interaction and "potential"

Functional Association Thru Interactions

- **Direct functional association:**
  - Interaction partners of a protein are likely to share functions w/ it
  - Proteins from the same pathways are likely to interact
- **Indirect functional association**
  - Proteins that share interaction partners with a protein may also likely to share functions w/ it
  - Proteins that have common biochemical, physical properties and/or subcellular localization are likely to bind to the same proteins

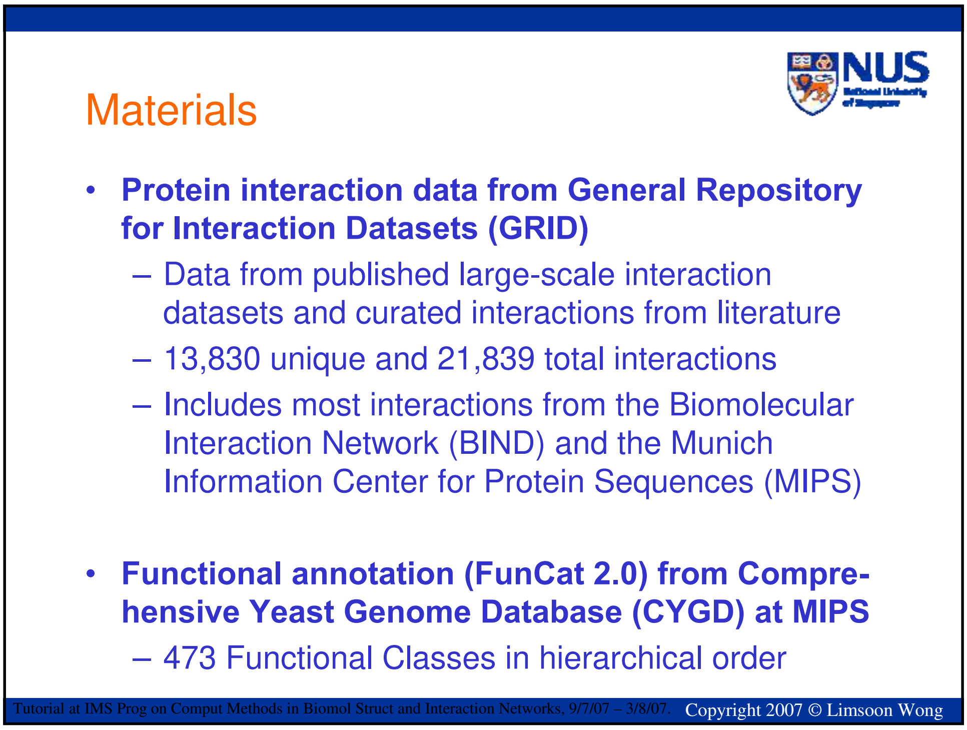






An Illustrative Case of Indirect Functional Association?

- **Is indirect functional association plausible?**
- **Is it found often in real interaction data?**
- **Can it be used to improve protein function prediction from protein interaction data?**

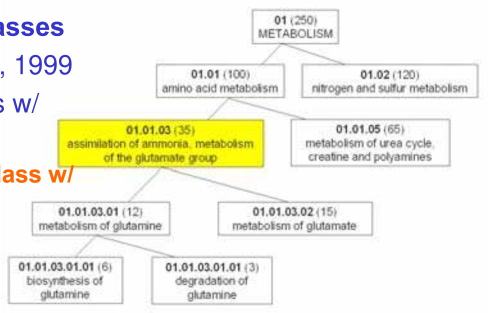


Materials

- **Protein interaction data from General Repository for Interaction Datasets (GRID)**
  - Data from published large-scale interaction datasets and curated interactions from literature
  - 13,830 unique and 21,839 total interactions
  - Includes most interactions from the Biomolecular Interaction Network (BIND) and the Munich Information Center for Protein Sequences (MIPS)
- **Functional annotation (FunCat 2.0) from Comprehensive Yeast Genome Database (CYGD) at MIPS**
  - 473 Functional Classes in hierarchical order

Validation Methods

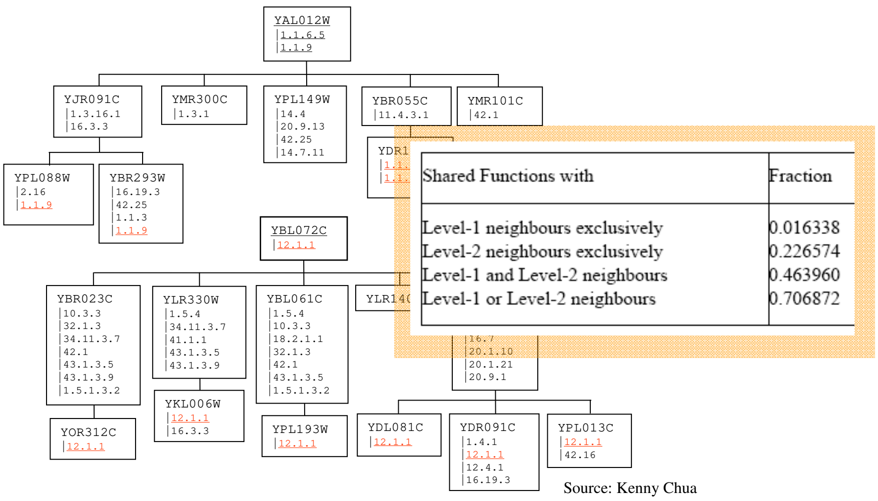
- **Informative Functional Classes**
  - Adopted from Zhou et al, 1999
  - Select functional classes w/
    - **at least 30 members**
    - **no child functional class w/ at least 30 members**



01 (250) METABOLISM
01.01 (100) amino acid metabolism
01.02 (120) nitrogen and sulfur metabolism
01.01.03 (35) assimilation of ammonia, metabolism of the glutamate group
01.01.05 (65) metabolism of urea cycle, creatine and polyamines
01.01.03.01 (12) metabolism of glutamine
01.01.03.02 (15) metabolism of glutamate
01.01.03.01.01 (6) biosynthesis of glutamine
01.01.03.01.01 (3) degradation of glutamine

- **Leave-One-Out Cross Validation**
  - Each protein with annotated function is predicted using all other proteins in the dataset

Freq of Indirect Functional Association



Shared Functions with	Fraction
Level-1 neighbours exclusively	0.016338
Level-2 neighbours exclusively	0.226574
Level-1 and Level-2 neighbours	0.463960
Level-1 or Level-2 neighbours	0.706872

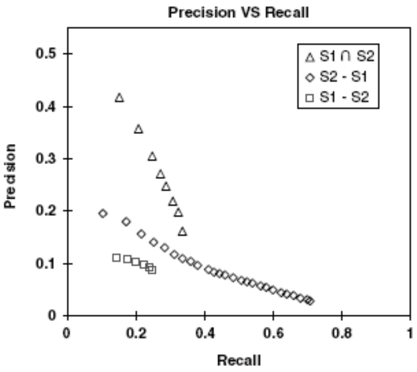
Source: Kenny Chua

Prediction Power By Majority Voting

- Remove overlaps in level-1 and level-2 neighbours to study predictive power of "level-1 only" and "level-2 only" neighbours
- Sensitivity vs Precision analysis

$$PR = \frac{\sum_i^K k_i}{\sum_i^K m_i} \qquad SN = \frac{\sum_i^K k_i}{\sum_i^K n_i}$$

- n_i is no. of fn of protein i
- m_i is no. of fn predicted for protein i
- k_i is no. of fn predicted correctly for protein i



⇒ "level-2 only" neighbours performs better

⇒ L1 ∩ L2 neighbours has greatest prediction power

Functional Similarity Estimate: 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**

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

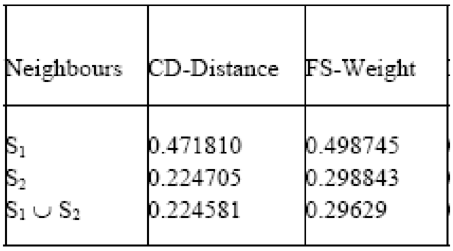
  - **N_k is the set of interacting partners of k**
  - **Greater weight given to similarity**

⇒ **Rewriting this as**

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

Correlation w/ Functional Similarity

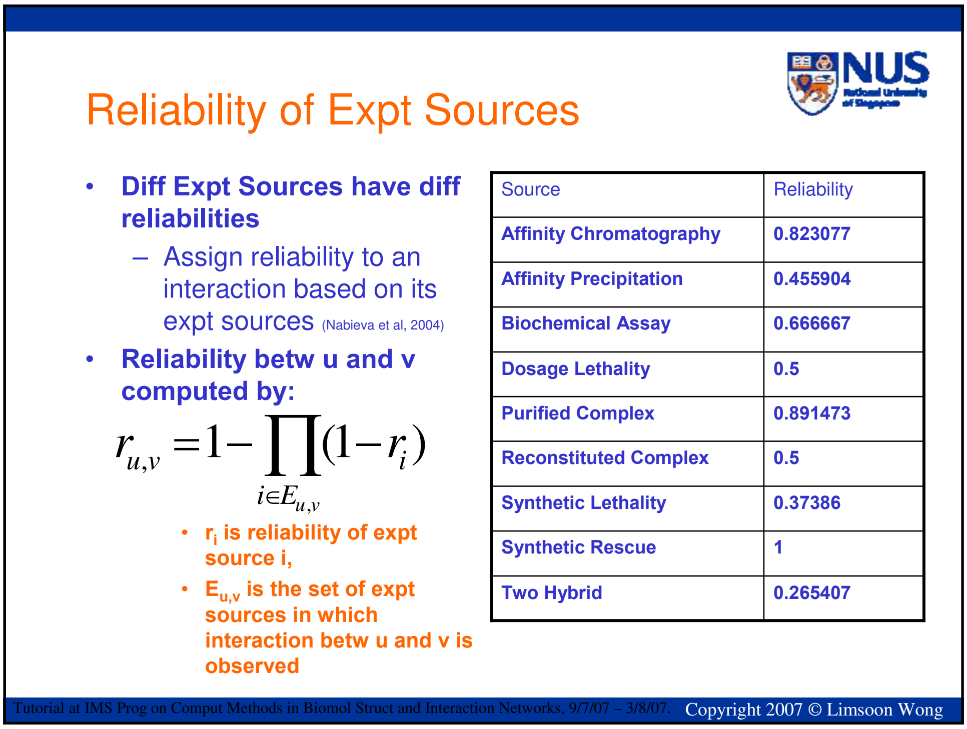
- **Correlation betw functional similarity & estimates**



Neighbours	CD-Distance	FS-Weight
S_1	0.471810	0.498745
S_2	0.224705	0.298843
$S_1 \cup S_2$	0.224581	0.29629

- **Equiv measure slightly better in correlation w/ similarity for L1 & L2 neighbours**

Source: Kenny Chua



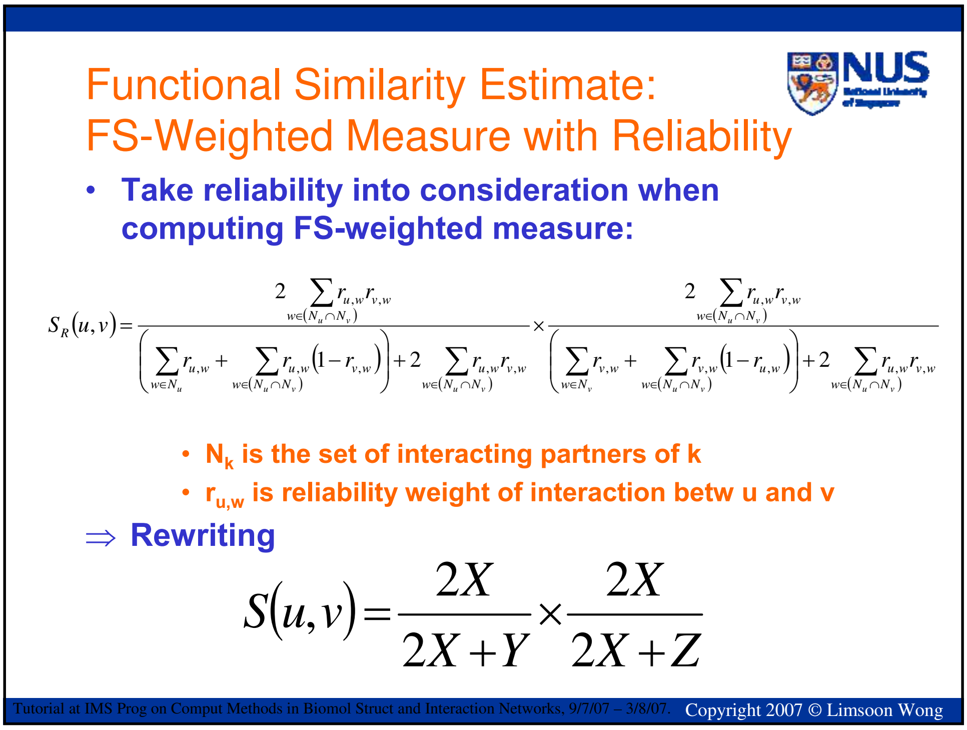
Reliability of Expt Sources

- **Diff Expt Sources have diff reliabilities**
  - Assign reliability to an interaction based on its expt sources (Nabieva et al, 2004)
- **Reliability betw u and v computed by:**

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

- **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**

Source	Reliability
Affinity Chromatography	**0.823077**
Affinity Precipitation	**0.455904**
Biochemical Assay	**0.666667**
Dosage Lethality	**0.5**
Purified Complex	**0.891473**
Reconstituted Complex	**0.5**
Synthetic Lethality	**0.37386**
Synthetic Rescue	**1**
Two Hybrid	**0.265407**



Functional Similarity Estimate: FS-Weighted Measure with Reliability

- **Take reliability into consideration when computing FS-weighted measure:**

$$S_R(u,v) = \frac{2 \sum_{w \in (N_u \cap N_v)} r_{u,w} r_{v,w}}{\left(\sum_{w \in N_u} r_{u,w} + \sum_{w \in (N_u \cap N_v)} r_{u,w} (1 - r_{v,w}) \right) + 2 \sum_{w \in (N_u \cap N_v)} r_{u,w} r_{v,w}} \times \frac{2 \sum_{w \in (N_u \cap N_v)} r_{u,w} r_{v,w}}{\left(\sum_{w \in N_v} r_{v,w} + \sum_{w \in (N_u \cap N_v)} r_{v,w} (1 - r_{u,w}) \right) + 2 \sum_{w \in (N_u \cap N_v)} r_{u,w} r_{v,w}}$$

- **N_k is the set of interacting partners of k**
- **$r_{u,w}$ is reliability weight of interaction betw u and v**

⇒ **Rewriting**

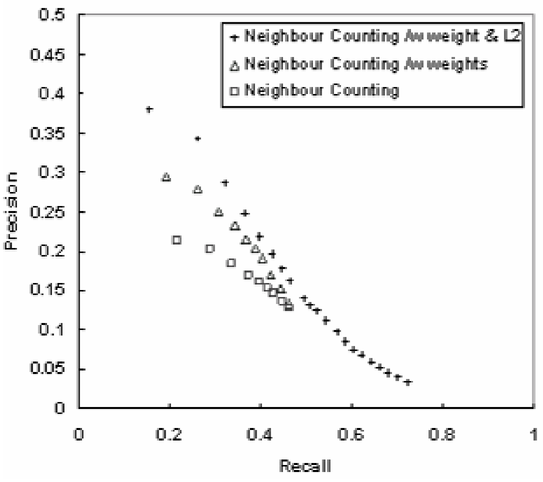
$$S(u,v) = \frac{2X}{2X+Y} \times \frac{2X}{2X+Z}$$

Integrating Reliability

- **Equiv measure shows improved correlation w/ functional similarity when reliability of interactions is considered:**

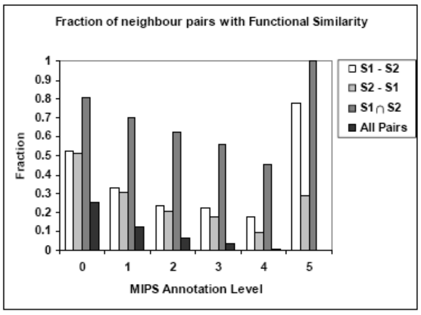
Neighbours	CD-Distance	FS-Weight	FS-Weight R
S_1	0.471810	0.498745	0.532596
S_2	0.224705	0.298843	0.375317
$S_1 \cup S_2$	0.224581	0.29629	0.363025

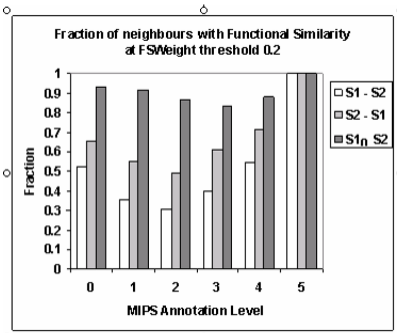
Improvement to Prediction Power by Majority Voting



Considering only neighbours w/ FS weight > 0.2

Improvement to Over-Rep of Functions in Neighbours





Source: Kenny Chua

Use L1 & L2 Neighbours for Prediction

- **FS-weighted Average**

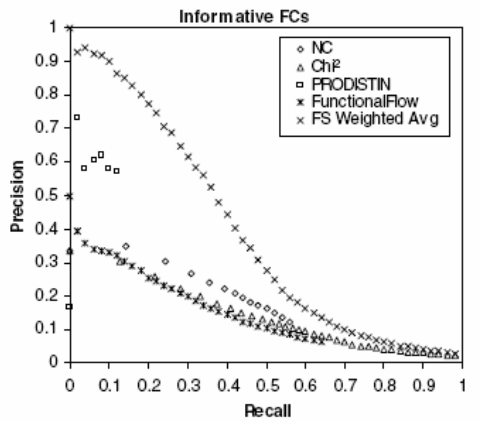
$$f_x(u) = \frac{1}{Z}\left[\lambda r_{\text{int}}\pi_x + \sum_{v\in N_u}\left(S_{TR}(u,v)\delta(v,x) + \sum_{w\in N_v} S_{TR}(u,w)\delta(w,x)\right)\right]$$

- **r_{int} is fraction of all interaction pairs sharing function**
- **λ is weight of contribution of background freq**
- **δ(k, x) = 1 if k has function x, 0 otherwise**
- **N_k is the set of interacting partners of k**
- **π_x is freq of function x in the dataset**
- **Z is sum of all weights**

$$Z = 1 + \sum_{v\in N_u}\left(S_{TR}(u,v) + \sum_{w\in N_v} S_{TR}(u,w)\right)$$

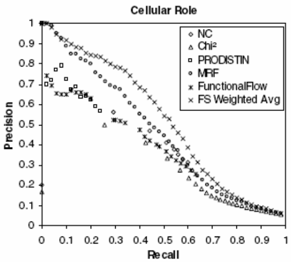
Performance of FS-Weighted Averaging

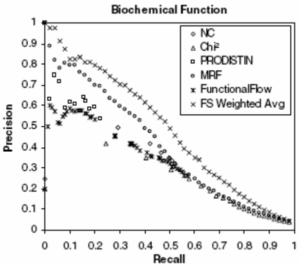
- **LOOCV comparison with Neighbour Counting, Chi-Square, PRODISTIN**

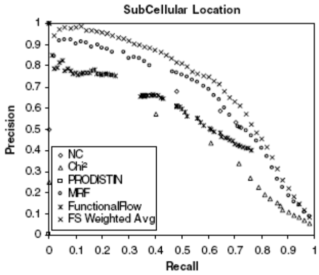


Performance of FS-Weighted Averaging

- **Dataset from Deng et al, 2003**
  - Gene Ontology (GO) Annotations
  - MIPS interaction dataset
- **Comparison w/ Neighbour Counting, Chi-Square, PRODISTIN, Markov Random Field, FunctionalFlow**





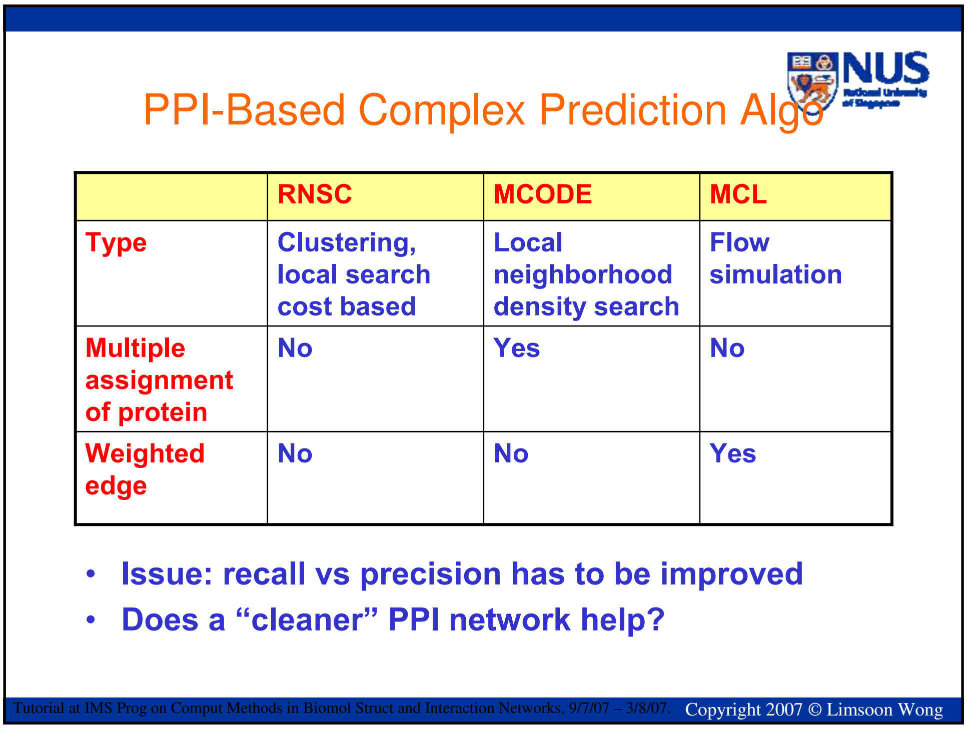


Conclusions

- Indirect functional association is plausible
- It is found often in real interaction data
- It can be used to improve protein function prediction from protein interaction data
- It should be possible to incorporate interaction networks extracted by literature in the inference process within our framework for good benefit

Another thing that we can use a more reliable PPI for:

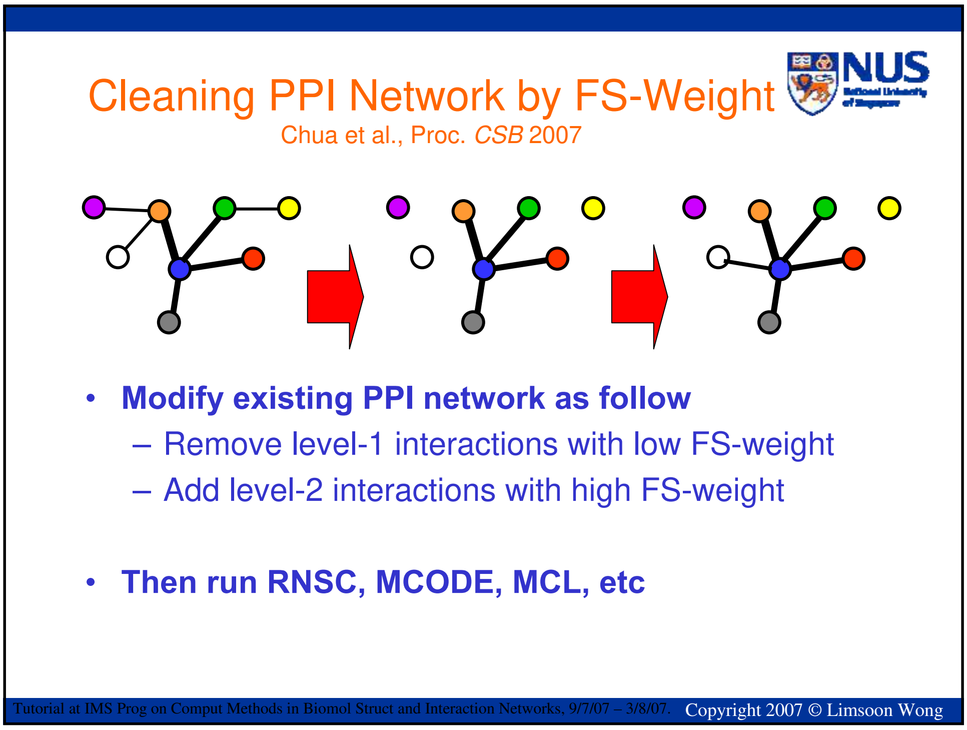
Protein complex prediction



PPI-Based Complex Prediction Algo

	RNSC	MCODE	MCL
Type	Clustering, local search cost based	Local neighborhood density search	Flow simulation
Multiple assignment of protein	No	Yes	No
Weighted edge	No	No	Yes

- Issue: recall vs precision has to be improved
- Does a “cleaner” PPI network help?



Cleaning PPI Network by FS-Weight

Chua et al., Proc. *CSB* 2007

- Modify existing PPI network as follow
  - Remove level-1 interactions with low FS-weight
  - Add level-2 interactions with high FS-weight
- Then run RNSC, MCODE, MCL, etc

Experiments

- **PPI datasets**
  - PPI[BioGRID], BioGRID db from Stark et al., 2006
- **Gold standards**
  - PC_{2004}, Protein complexes from MIPS 03/30/2004
  - PC_{2006}, Protein complexes from MIPS 05/18/2006

- **Validation criteria**

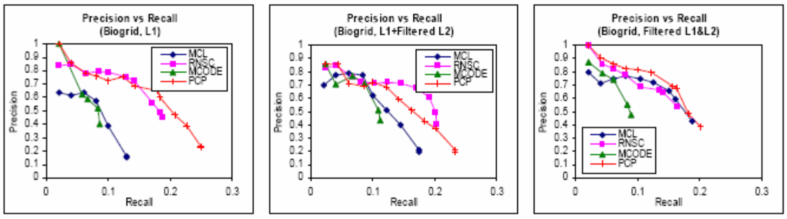
$$overlap(S,C) = \frac{|V_S \cap V_C|^2}{|V_S| \cdot |V_C|}$$

where

- S = predicted cluster
- C = true complex
- V_x = vertices of subgraph defined by X

- **Overlap(S,C) ≥ 0.25 is considered a correct prediction**

Validation on PC_{2004}

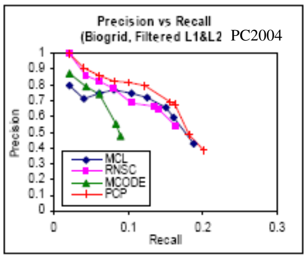


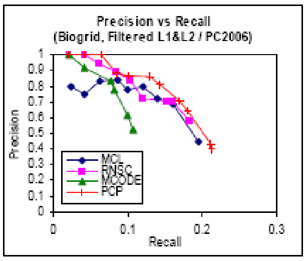
(d) (e) (f)

a) *Original level-1 PPI*
b) *Original level-1 PPI and filtered level-2 PPI*
c) *Filtered level-1 and level-2 PPI*

- **Precision is improved in all methods**
- **PCP (more later) performs best**

Validation on PC_{2006}





- **When predictions are validated against PC_{2006}, precision of all algo improved**
- **Many "false positives" wrt PC_{2004} are actually real**
- **PCP again performs best**

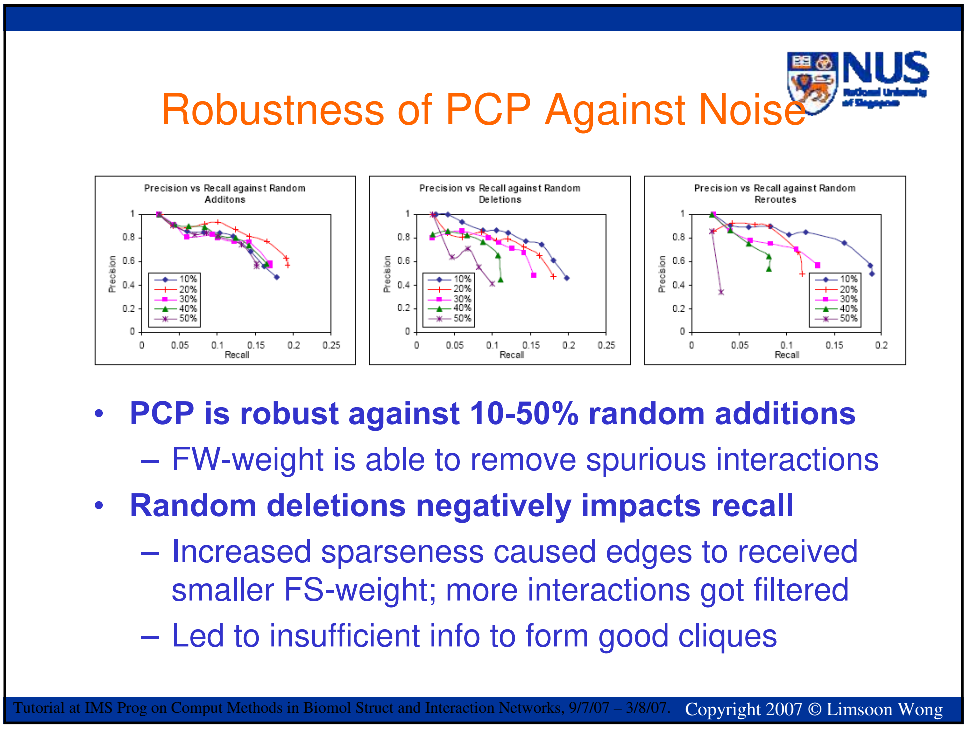
PCP Algorithm

Chua et al., Proc. *CSB* 2007

- **Find all max cliques in the modified PPI network**
  - If two cliques overlap, distribute the overlapped nodes such that both cliques have larger average FS-weight
- **Merge resulting (partial) cliques with good inter-cluster density**

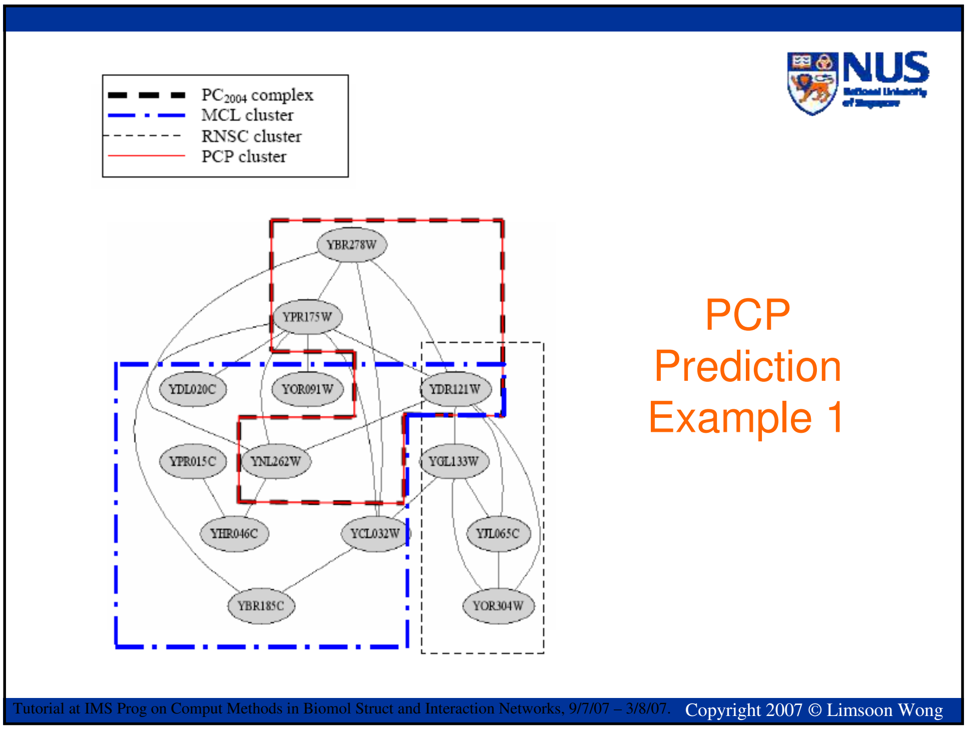
$$ICD(S_a, S_b) = \frac{\sum S_{FS}(i,j) \mid i \in (V_a - V_b), j \in (V_b - V_a), (i,j) \in E}{|V_a - V_b| \cdot |V_b - V_a|}$$

- **Modify the PPI network by treating the merged partial cliques as vertices**
- **Iterate the steps above**

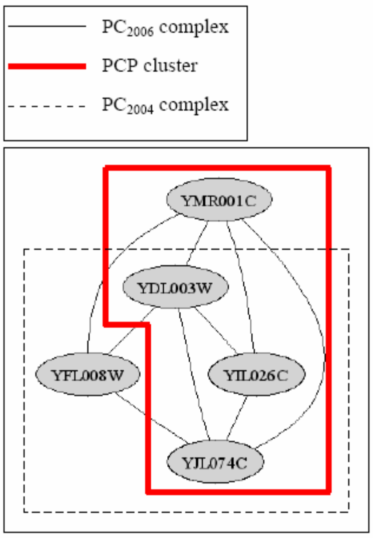


Robustness of PCP Against Noise

- **PCP is robust against 10-50% random additions**
  - – FW-weight is able to remove spurious interactions
- **Random deletions negatively impacts recall**
  - – Increased sparseness caused edges to received smaller FS-weight; more interactions got filtered
  - – Led to insufficient info to form good cliques



PCP Prediction Example 1



PCP Prediction Example 2

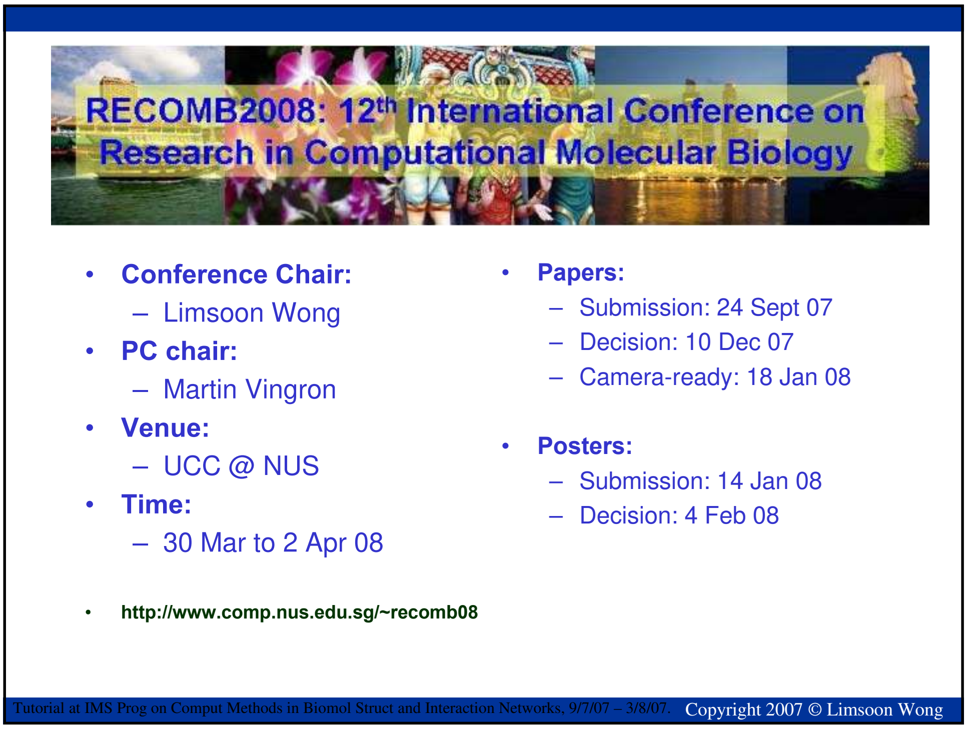
Conclusions

- **Precision of protein complex prediction can be improved by**
  - PPI network augmented with level-2 interactions
  - PPI network cleansed by FS-weight
- **PCP performs excellently**

References

- **J. Chen et al,** "Increasing confidence of protein-protein interactomes", **Proc. GIW 2006, pages 284—297**
- **J. Chen et al,** "Systematic assessment of high-throughput protein interaction data using alternative path approach", ***Proc. ICTAI 2004*, pages 368—372**
- **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 et al.** "Using indirect protein-protein interactions for protein complex prediction", **Proc. CSB 2007, August 2007, to appear**
- **R. Saito et al,** "Interaction generality, a measurement to assess the reliability of a protein-protein interaction", ***NAR* 30:1163--1168, 2002**
- **R. Saito et al,** "Construction of reliable protein-protein interaction networks with a new interaction generality measure", ***Bioinformatics* 19:756--763, 2003**

- **PC chairs:**
  - See-Kiong Ng
  - Hiroshi Mamitsuka
- **Venue:**
  - Biopolis @ One North
- **Time:**
  - 3 to 5 Dec 07
- **http://www.comp.nus.edu.sg/~giw2007**
- **Papers:**
  - Submission: 13 May 07
  - Decision: 15 Jul 07
  - Camera-ready: 5 Aug 07
- **Posters:**
  - Submission: 16 Sep 07
  - Decision: 14 Oct 2007



RECOMB2008: 12th International Conference on Research in Computational Molecular Biology

- **Conference Chair:**
  - Limsoon Wong
- **PC chair:**
  - Martin Vingron
- **Venue:**
  - UCC @ NUS
- **Time:**
  - 30 Mar to 2 Apr 08
- **Papers:**
  - Submission: 24 Sept 07
  - Decision: 10 Dec 07
  - Camera-ready: 18 Jan 08
- **Posters:**
  - Submission: 14 Jan 08
  - Decision: 4 Feb 08

- **http://www.comp.nus.edu.sg/~recomb08**