Workshop on Data Analysis and Data Mining in Proteomics

Detecting False Positives and False Negatives in Protein Interactome using Network Topology

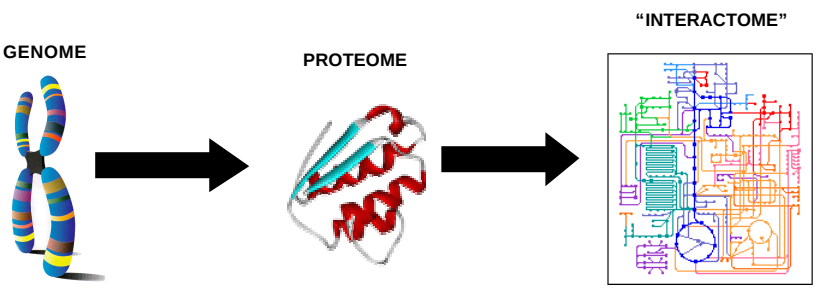
See-Kiong Ng

Knowledge Discovery Department, I²R

12 May 2005

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** with other proteins and biomolecules.



High Tech Experimental Methods

Many **high-throughput experimental detection methods** for protein-protein interactions have been devised; e.g. yeast two-hybrid assays.

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

Key Bottleneck:

High-throughput approach sacrifice quality for ***quantity***:

(a) limited or biased coverage: ***false negatives***, and

(b) high error rates : ***false positives***

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

Unravelling the Protein-Protein Interactome

Part I:

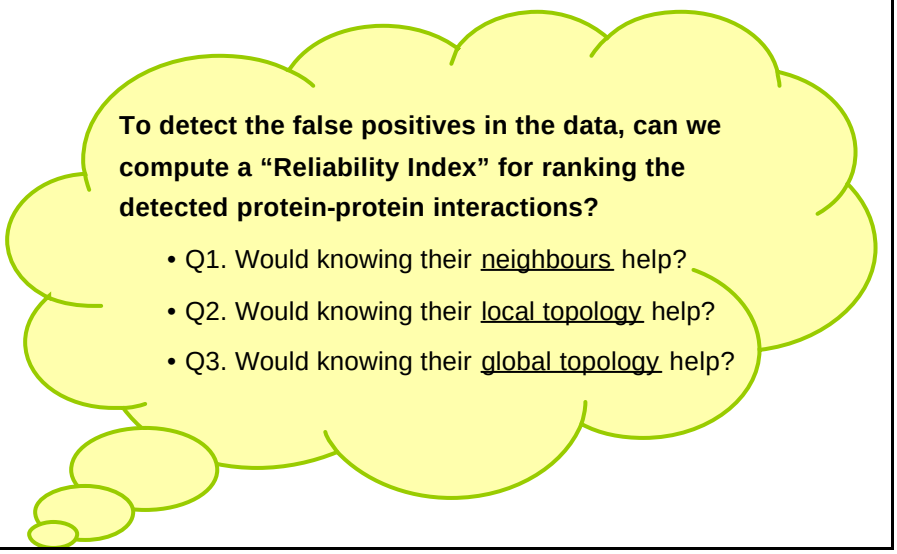
The False Positive Challenge

How reliable are experimentally-derived protein interactions?

Some high-throughput protein interaction experiments, such as the popular Y2H method, have as much as 50% false positives.

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

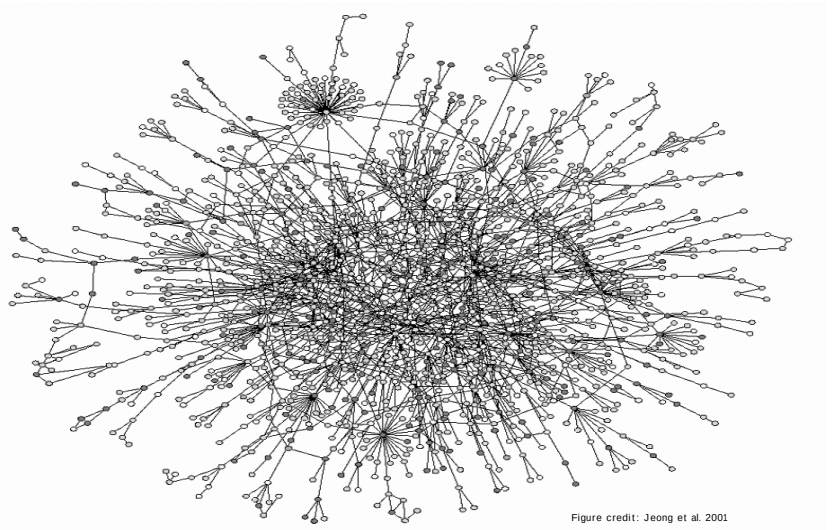
Creating a Reliability Index



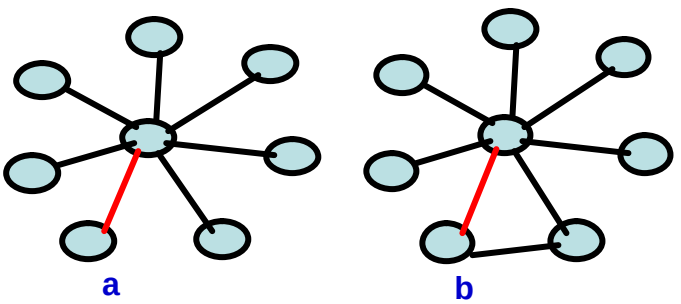
Part I: Would knowing their neighbours help?

The story of IG1

Real-World Network Model for PPI



An Observation



"It seems that configuration **a** is less likely than configuration **b** for protein interaction networks."

Why?

- In real-world, "**My friends are likely to know each other!"**
- But how to quantitate this observation?

Interaction Generality (IG1)

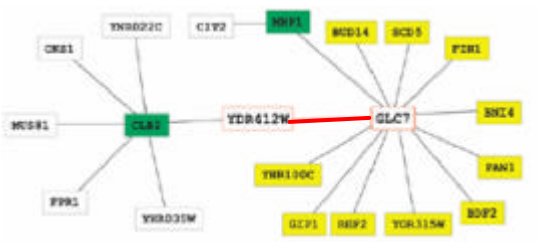
Saito et al., *NAR*, 30:1163-1168, 2002

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 $\mathcal{G}$.

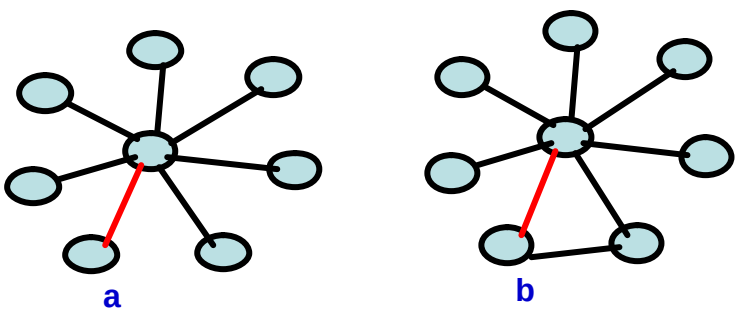
The number of proteins that "interact" with just X or Y, and nobody else



The **lower** the IG value, the more reliable

Interaction Generality IG1

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



a **b**

$IG_a = 6$ < $IG_b = 5$

According to their IG values, the **red** interaction is more likely to be a true interaction in configuration **b** than in configuration **a**.

Using IG1 to Evaluate Protein Interactions

I.G.	Ito ol.	ovlap	Uetz ol.	ovlap
1	229	66	236	58
2	137	34	226	37
3	57	16	113	16
4	43	6	66	6
5	24	4	38	5
6	16	1	37	2
7	27	0	20	3
8	23	1	16	2
9	9	1	4	0
10	2	0	44	0
11	0	0	9	2
12	1	0	4	0
13	13	0	0	1
14	15	0	1	1
15	16	0	0	0
16	30	3	1	0
17	6	1	0	0
18	20	0	0	0
19	2	0	0	0
20	3	0	0	0
21	0	0	0	0
22	0	0	0	0
23	0	0	0	0
24	0	0	0	0
25	0	0	0	0
26–	0	0	0	0
Total	673	133	815	133

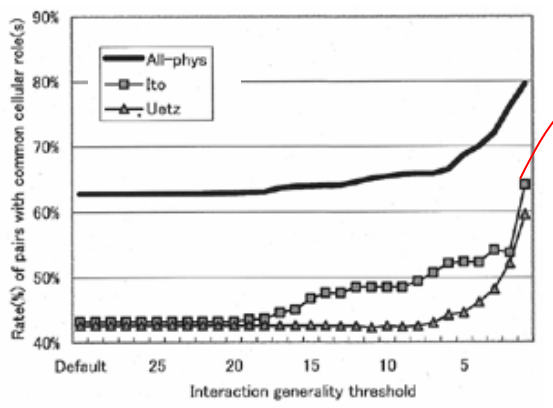
There are 229 pairs in Ito having ig = 1.
Of these, 66 are also reported by Uetz

- Interacting pairs common to Ito *et al.* & Uetz *et al.* are more reliable
- These also have smaller IG1 values

⇒ **IG1 seems to be able to pick out the more reliable interactions**

Using IG1 to Evaluate Protein Interactions

~65% of pairs in in Ito having ig=1 are known to have common localization



- Interaction pairs having common cellular localization are more likely to be real
- These also have lower IG1 values

⇒ **IG1 seems to work**

Part II: Would knowing their local topology help?

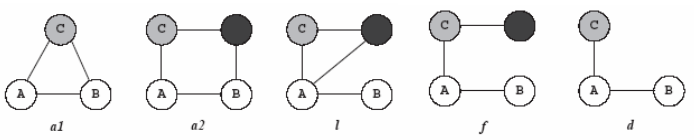
The story of IG2

Looking Beyond Immediate Neighbours

Saito et al., *Bioinformatics*, 19:756--763, 2002

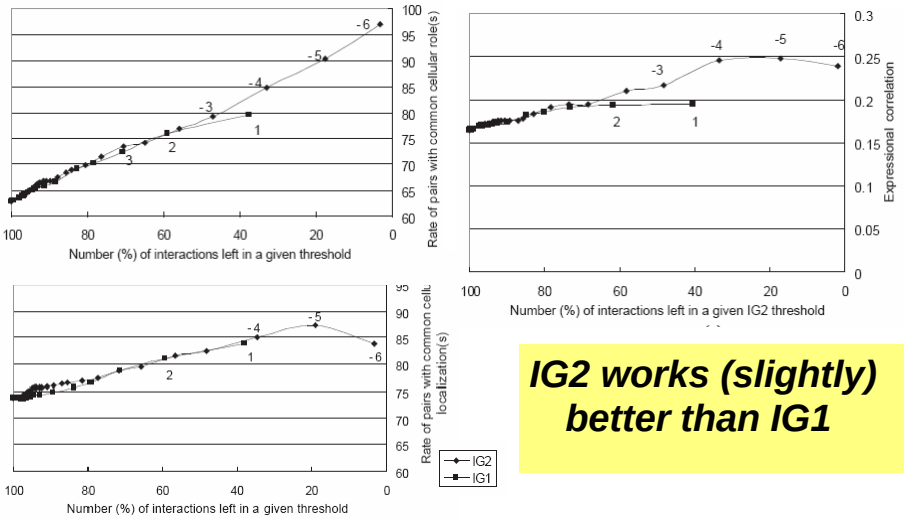
IG2

- A new interaction generality measure based on the 5 possible local topological configurations of an interaction **A-B** w.r.t. a third protein **C**



- IG2 was found to perform better than IG1, showing that a **more global network topological** approach is advantageous

IG2 vs IG1



IG2 works (slightly) better than IG1

Saito et al., Bioinformatics, 19:756--763, 2002

Part III: Would knowing their global topology help?

The story of IRAP

Interaction Pathway Reliability

Conjecture:

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

- Biological functions are typically performed by highly interconnected networks of interactions
- Circular contigs are frequently observed in protein interaction and complex data
- Alternate paths are also observed in many real-world networks

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

PPI Network Model

- Model PPI as a *weighted* undirected network *G=(V, E)* where
  - Each node in *G* denotes a protein
  - An edge from node v_A to node v_B denotes an interaction between protein A and protein B
  - Initial edge weight:

$$weight(v_A, v_B) = 1 - \frac{IG1(A,B)}{IG1_{max}}$$

Interaction Pathway Reliability

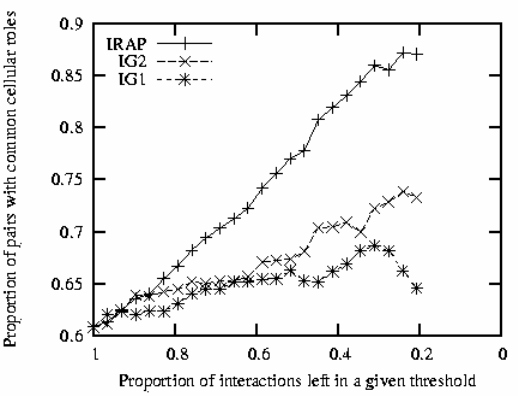
IRAP : Interaction Reliability by Alternate Pathway

A novel **global topology** measure for assessing interaction reliability computationally

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

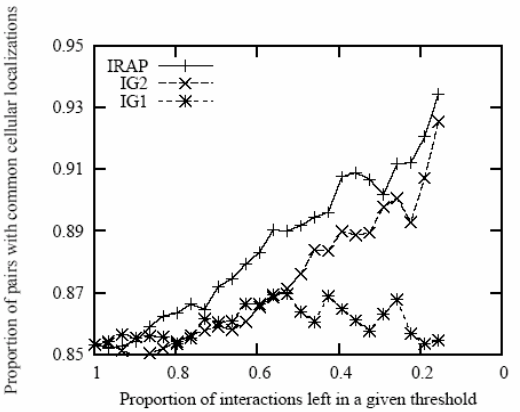
where $ig^{\mathcal{G}}_{\max} = \max\{ig^{\mathcal{G}}(X \leftrightarrow Y) \mid (X \leftrightarrow Y) \in \mathcal{G}\}$ is the maximum *IG1* value in *G*, and $\Phi^{\mathcal{G}}(X, Y)$ is the set of all possible non-reproducible paths between *X* and *Y*, but excluding the direct path *X « Y.*

Evaluation of IRAP: Cellular Functions



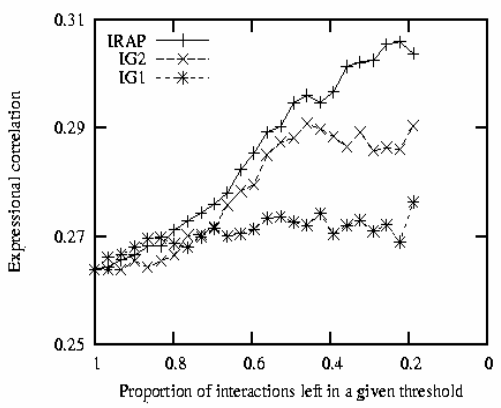
IRAP outperforms both IG1 and IG2 in terms of the proportion of interactions with common cellular roles.

Evaluation of IRAP: Co-Localization



IRAP outperforms both IG1 and IG2 in terms of the proportion of co-localized interactions.

Evaluation of IRAP: Co-Expression



IRAP outperforms both IG1 and IG2 in terms of the proportion of genetically co-expressed interactions.

Discovering Biologically Interacting "Cross-Talkers"

- From our co-localization experiment, we also observed that there are 257 **non-co-localized protein pairs with high IRAP** (>0.95)
- A large proportion (53%) of these cross-talking pairs have the same MIPS functional annotation, suggesting that they could be ***biologically interacting cross-talkers*** such as those in signal tranduction pathways

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

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

Unravelling the Protein-Protein Interactome

Part II:

The False Negative Challenge

Approach

- Use the Alternative Path Model to detect false negatives in an experimental PPI network:
  - Compute an IRAP value for each *unlinked* protein pair in the network
  - If it has a high IRAP value, then it may be a false negative

Experiment

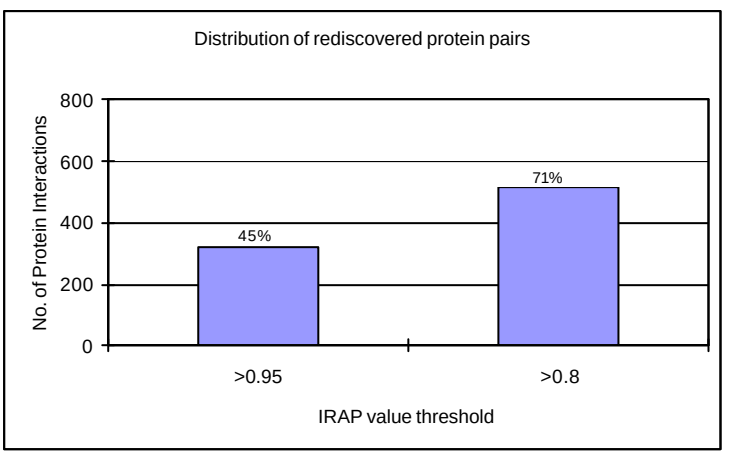
Dataset

- 10,199 non-redundant interactions between 4,336 yeast proteins from MIPS with date Jan. 18, 2005
- "Verified true" interactions in PPI network
  - Ito. core set (833 interactions)

- Apply IRAP on dataset with the 833 true interactions hidden from the program

- IRAP program re-discovered 730 interactions

Results



Re-discovered true interactions have high IRAP values.

Results

- Compare with random connections

Subcellular Localization		Protein Pairs									
		YDL130W	YLR287C	YPR173C	YOR276C	YDR259C	YDR311W	YGL163W	YOR180C	YOR327C	YDL226C
725	YDL130W		0.97403			0.97403					
	YLR287C										
726/750	YPR173C				0.97419						
	YOR276C										
750	YDR259C						0.97403	0.96154		0.98701	0.97403
	YDR311W							0.96138		0.96138	
750	YGL163W								0.98701	0.97419	0.96138
	YOR180C										
750	YOR327C										0.96138
	YDL226C										

Only values greater than 0.95 are shown in the table

A Case Study

Yeast 26S Proteosome Complex

- 4 sub-structures found in PDB covering **13 proteins**
- **23 physical interactions** using the (10, 5)-rule

- Out of the 23 physical interactions in this complex, only **1** was detected in the MIPS experimental dataset

- IRAP detected 42 interactions (out of 78 possible interactions) between the 13 proteins:
  - **14** of these are the physical interactions in the complex (c.f. 1 detected in MIPS)
  - Another 9 were experimentally detected interactions in MIPS
  - 19 new interactions were also predicted

Conclusions

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

Acknowledgements

- **Chen Jin**, PhD Student, SoC, NUS

- A/Prof. Wynne Hsu, SoC, NUS

- A/Prof. Mongli Lee, SoC, NUS

- Prof. Limsoon Wong, I^2R

References

J. Chen, W. Hsu, M.L. Lee, S.-K. Ng (2004) "**Systematic Assessment of High-Throughput Experimental Data for Reliable Protein Interactions using Network Topology**", in Proceedings of the *16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004)*, November 15-17, Florida, pages 368-372.

J. Chen, W. Hsu, M.L. Lee, and S.-K. Ng (2005) "**Discovering Reliable Protein Interactions from High-Throughput Experimental Data using Network Topology**", accepted for publication in *Artificial Intelligence in Medicine.*

Workshop on Data Analysis and Data Mining in Proteomics

THANK YOU