Challenges in Understanding Pathways, Predicting Complexes, & Inferring Protein Function

Limsoon Wong

Plan

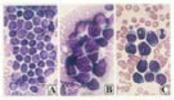
- **Understanding Pathways**
  - Past successes
  - Towards more meaningful genes
  - Issues on pathway sources
- **Predicting Complexes**
  - Current approaches
  - Issues on network noise and density assumption
  - Benefits of network cleansing
- **Inferring Protein Function**
  - Guilt-by-association
  - When guilt-by-association fails

Understanding Pathways

Childhood Acute Lymphoblastic Leukemia

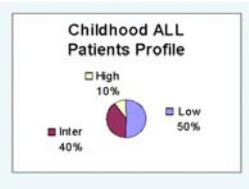
- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse

- **The subtypes look similar**



- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
  - ⇒Unavailable in developing countries

Patient Profiles & Treatment Costs



- **2000 new cases a year in ASEAN countries**

- **Treatment for childhood ALL over 2 yrs**
  - Intermediate intensity: US$60k
  - Low intensity: US$36k
  - High intensity: US$72k
- **Treatment for relapse: US$150k**
- **Cost for side-effects: Unquantified**

Why not high/low intensity to everyone?

- **High-intensity Tx**
  - Over intensive for 90% of patients, thus a lot more side effects
  - US$144m (US$72k * 2000) for high-intensity tx

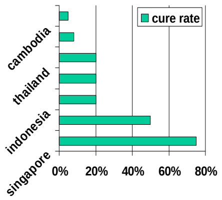
⇒ **Total US$144m/yr plus un-quantified costs for dealing with side effects**

- **Low-intensity Tx**
  - Under intensive for 50% of patients, thus a lot more relapse
  - US$72m (US$36k * 2000) for low-intensity tx
  - US$150m (US$150k * 2000 * 50%) for relapse tx

⇒ **Total US$222m/yr**

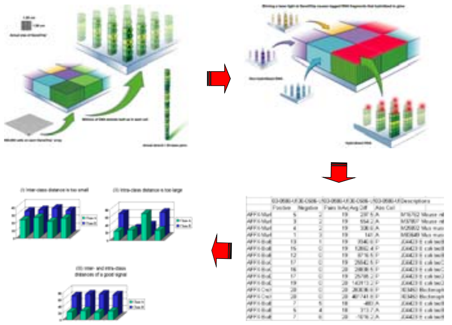
Current Situation

- Intermediate intensity conventionally applied in ASEAN countries



- Over intensive for 50% of patients, thus more side effects
- Under intensive for 10% of patients, thus more relapse
- US$120m (US$60k * 2000) for intermediate intensity tx
- US$30m (US$150k * 2000 * 10%) for relapse tx
- Total US$150m/yr plus un-quantified costs for dealing with side effects

Single-Test Platform of Microarray & Machine Learning



Individual Gene Testing

Fold Change

$$FC_{ratio_i} = \frac{x_i}{y_i} \qquad FC_{diff_i} = x_i - y_i$$

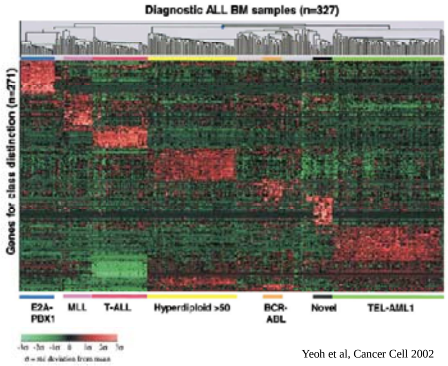
x – Microarray value after drug
y – Microarray value before drug
i – Gene

T-test

$$T_i = \frac{\hat{x}_i - \hat{y}_i}{s_i} \qquad T_i = \frac{\hat{x}_i - \hat{y}_i}{s_i + s_o} \qquad T_i = \frac{\hat{x}_i - \hat{y}_i}{\sqrt{Bs^2 + (1-B)s_i^2}}$$

x – Log2 value of treatment
y – Log2 value of control
s – Standard error
i – Gene

Golub et al, Science, 1999



Yeoh et al, Cancer Cell 2002

Exploit Invariant Gene Expr Profiles

- Low intensity applied to 50% of patients
- Intermediate intensity to 40% of patients
- High intensity to 10% of patients

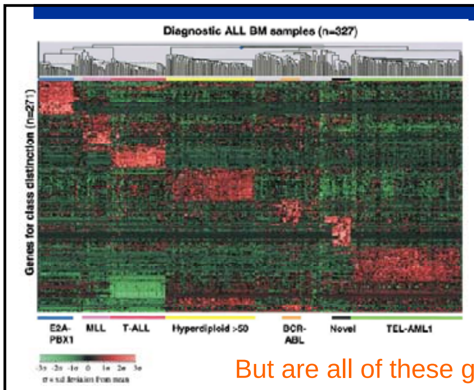
⇒ Reduced side effects
⇒ Reduced relapse
⇒ 75-80% cure rates

- US$36m (US$36k * 2000 * 50%) for low intensity
- US$48m (US$60k * 2000 * 40%) for intermediate intensity
- US$14.4m (US$72k * 2000 * 10%) for high intensity

- Total US$98.4m/yr

⇒ Save US$51.6m/yr

Yeoh et al, Cancer Cell 2002



But are all of these genes meaningful?

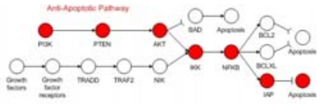
Percentage of Overlapping Genes

- **Low % of overlapping genes from diff expt in general**
  - Prostate cancer
    - **Lapointe et al, 2004**
    - **Singh et al, 2002**
  - Lung cancer
    - **Garber et al, 2001**
    - **Bhattacharjee et al, 2001**
  - DMD
    - **Haslett et al, 2002**
    - **Pescatori et al, 2007**

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

Zhang et al, Bioinformatics, 2009

Gene Regulatory Circuits



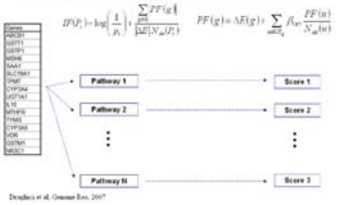
- **Each disease subtype has underlying cause**
- **There is a unifying biological theme for genes that are truly associated with a disease subtype**

- **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**

Towards More Meaningful Genes

- **ORA**
  - Khatri et al
  - Genomics, 2002
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - PNAS, 2005
- **Pathway Express**
  - Draghici et al
  - Genome Res, 2007

Gene Class Testing: Pathway Express

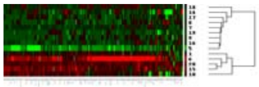


$$IF(P_i) = \log\left(\frac{1}{p_i}\right) + \frac{\sum_{g \in P_i} |PF(g)|}{|\Delta E| \cdot N_{de}(P_i)} \qquad PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \frac{PF(u)}{N_{ds}(u)}$$

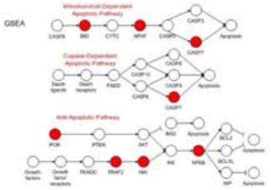
Draghici et al. Genome Res. 2007

Nasopharyngeal Carcinoma

- **NPC patients respond differentially to CYC202**



- **Can we identify drug action pathways by these more sophisticated methods?**

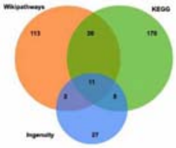


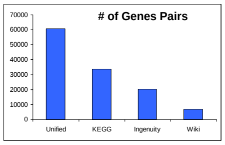
Futhermore,

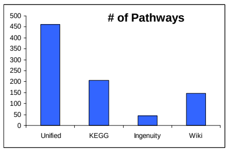
All of these newer methods rely on gene group or pathway information.

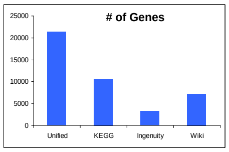
But how good are the available sources of pathway information?

Low Comprehensiveness of Pathway Sources

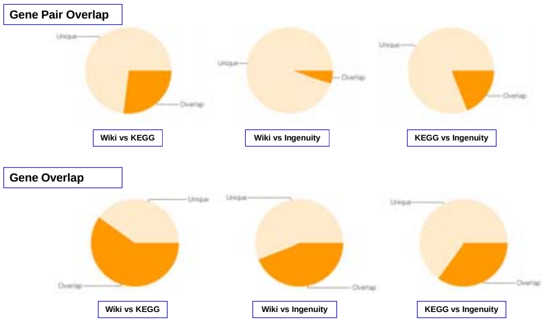








Low Consistency of Pathway Sources



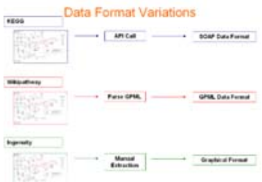
Example: Apoptosis Pathway

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%

BIOCARTA GenMAPP PharmGKB INGENUITY SYSTEMS

Would Unifying Pathway Sources Help?

- **Incompatibility Issues!**

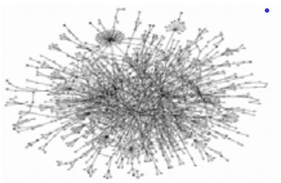


- **Data extraction method variations**
- **Format variations**
- **Data differences**
- **Pathway name differences**
- **Gene/GeneID name differences**

Recovering Space-Time Info from PPI Networks

Motivation

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

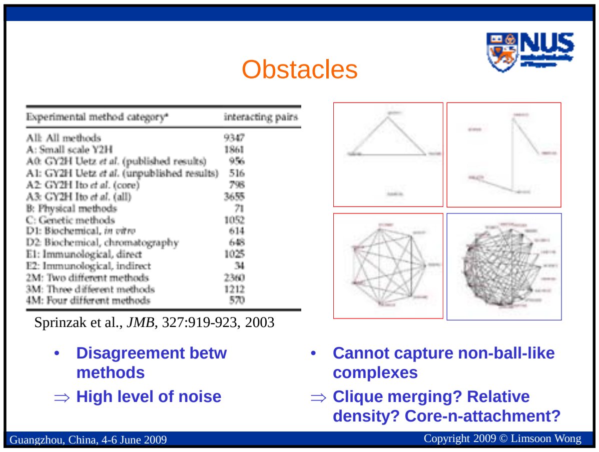


- **Can a protein interact with so many proteins simultaneously?**
- **A big “hub” and its “spokes” should probably be decomposed into subclusters**
  - Each subcluster is a set proteins that interact in the same space and time
  - Viz., a protein complex

Approaches to PPI-Based Protein Complex Prediction

	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

- **Recall vs precision is poor**
  - Noise in PPI network?
  - Non-ball-like complexes?



Obstacles

Experimental method category*	interacting pairs
All: All methods	9347
A: Small scale Y2H	1861
A0: GY2H Uetz *et al.* (published results)	956
A1: GY2H Uetz *et al.* (unpublished results)	516
A2: GY2H Ito *et al.* (core)	798
A3: GY2H Ito *et al.* (all)	3655
B: Physical methods	71
C: Genetic methods	1052
D1: Biochemical, *in vitro*	614
D2: Biochemical, chromatography	648
E1: Immunological, direct	1025
E2: Immunological, indirect	34
2M: Two different methods	2360
3M: Three different methods	1212
4M: Four different methods	570

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

- **Disagreement betw methods**

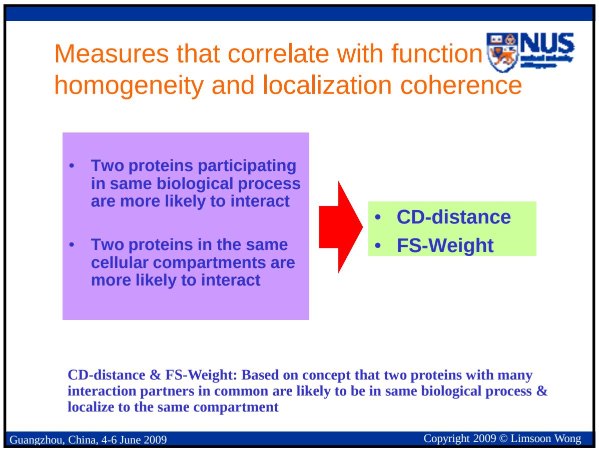
⇒ **High level of noise**

- **Cannot capture non-ball-like complexes**

⇒ **Clique merging? Relative density? Core-n-attachment?**

Guangzhou, China, 4-6 June 2009

Copyright 2009 © Limsoon Wong



Measures that correlate with function homogeneity and localization coherence

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

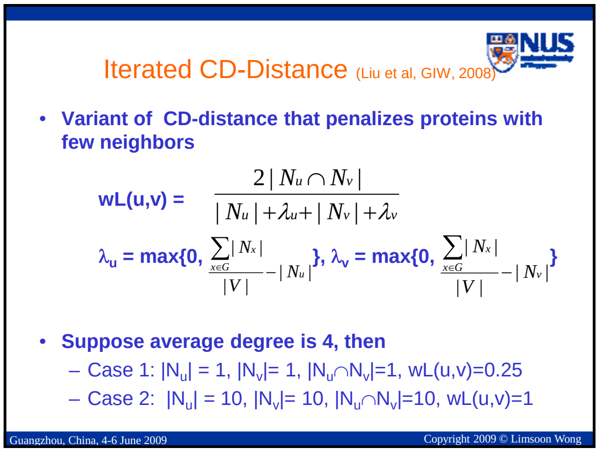
⇒

- **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

Guangzhou, China, 4-6 June 2009

Copyright 2009 © Limsoon Wong



Iterated CD-Distance (Liu et al, GIW, 2008)

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

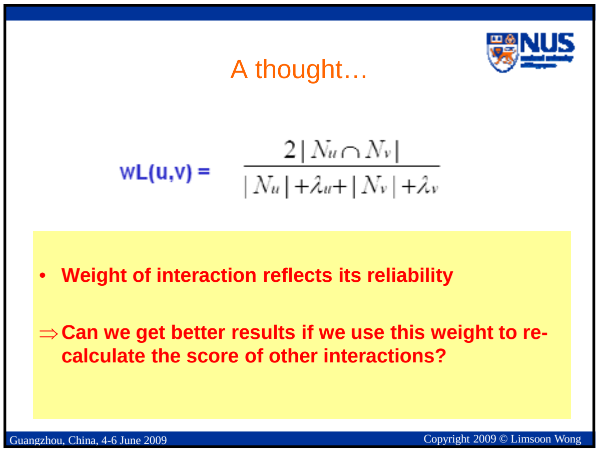
$$\textbf{wL(u,v)} = \frac{2\,|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

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

- **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: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, $wL(u,v) = 1$

Guangzhou, China, 4-6 June 2009

Copyright 2009 © Limsoon Wong



A thought…

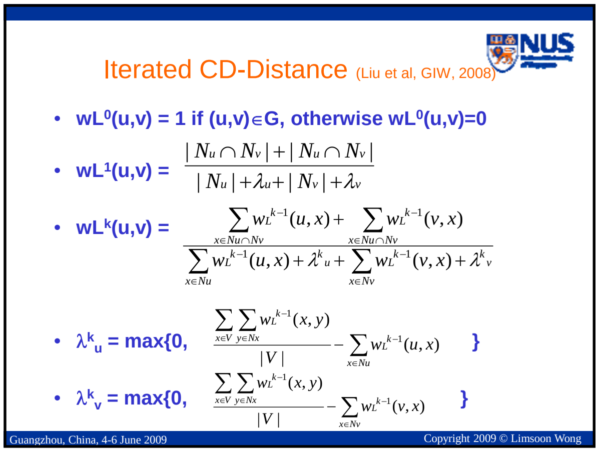
$$\textbf{wL(u,v)} = \frac{2\,|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

- **Weight of interaction reflects its reliability**

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

Guangzhou, China, 4-6 June 2009

Copyright 2009 © Limsoon Wong

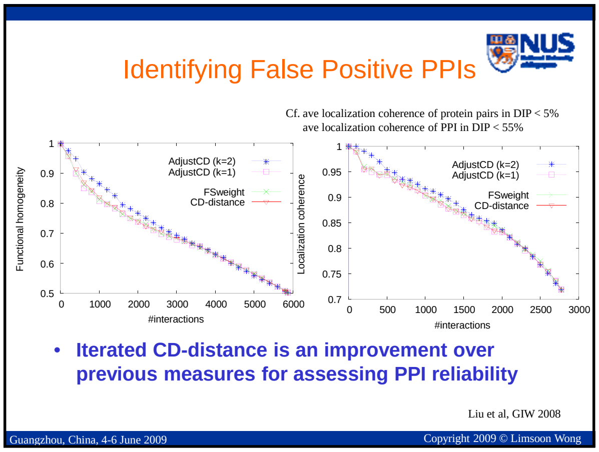


Iterated CD-Distance (Liu et al, GIW, 2008)

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

Guangzhou, China, 4-6 June 2009

Copyright 2009 © Limsoon Wong



Identifying False Positive PPIs

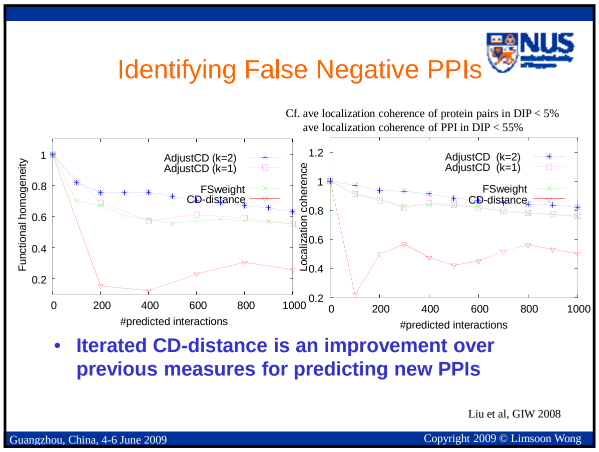
Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%

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

Liu et al, GIW 2008

Guangzhou, China, 4-6 June 2009

Copyright 2009 © Limsoon Wong

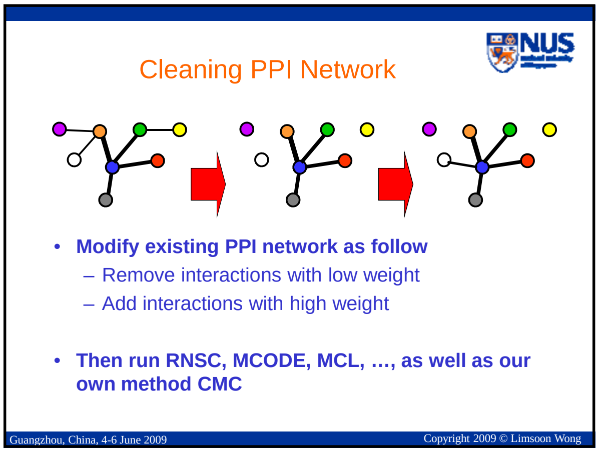


Identifying False Negative PPIs

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

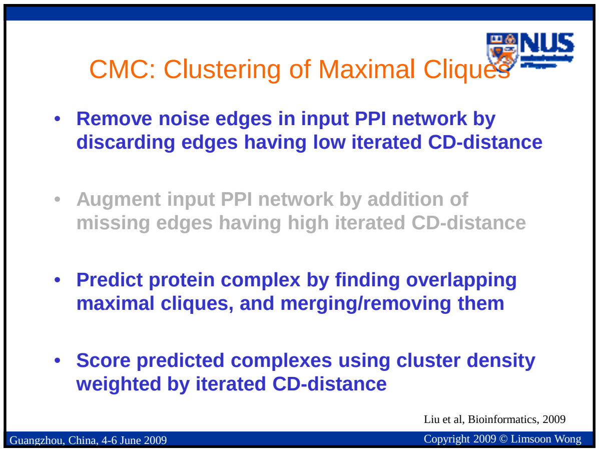
- Iterated CD-distance is an improvement over previous measures for predicting new PPIs

Liu et al, GIW 2008



Cleaning PPI Network

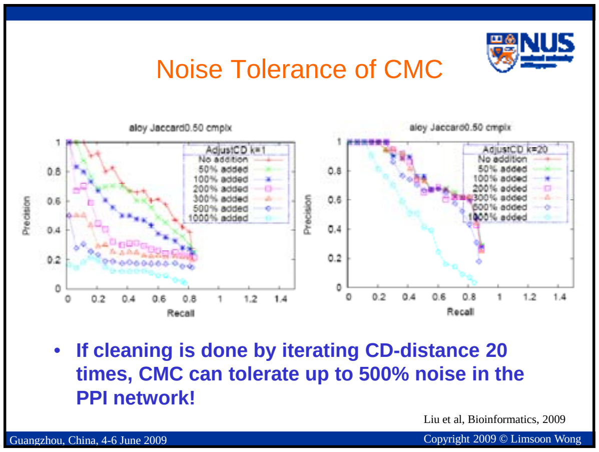
- Modify existing PPI network as follow
  - Remove interactions with low weight
  - Add interactions with high weight
- Then run RNSC, MCODE, MCL, …, as well as our own method CMC



CMC: Clustering of Maximal Cliques

- Remove noise edges in input PPI network by discarding edges having low iterated CD-distance
- Augment input PPI network by addition of missing edges having high iterated CD-distance
- Predict protein complex by finding overlapping maximal cliques, and merging/removing them
- Score predicted complexes using cluster density weighted by iterated CD-distance

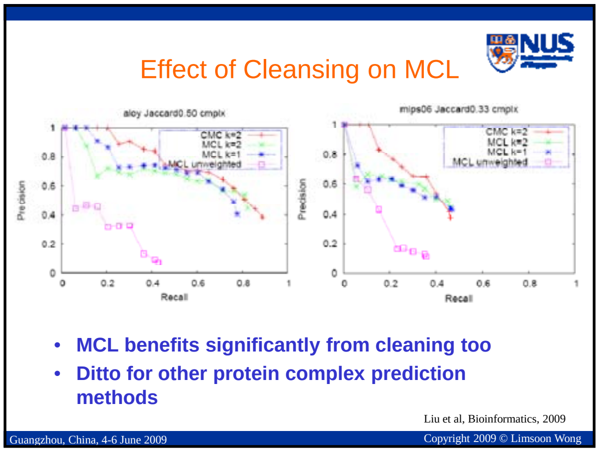
Liu et al, Bioinformatics, 2009



Noise Tolerance of CMC

- If cleaning is done by iterating CD-distance 20 times, CMC can tolerate up to 500% noise in the PPI network!

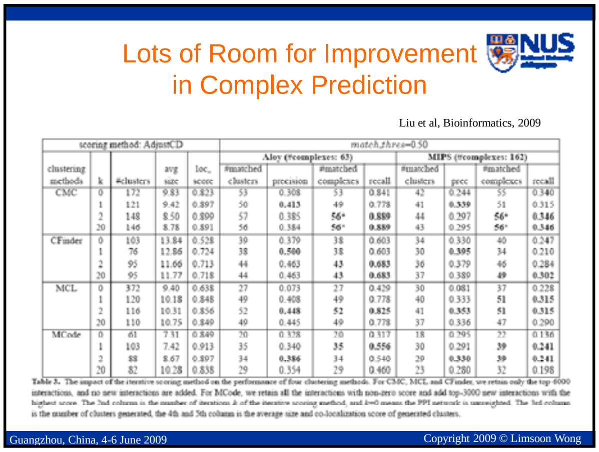
Liu et al, Bioinformatics, 2009



Effect of Cleansing on MCL

- MCL benefits significantly from cleaning too
- Ditto for other protein complex prediction methods

Liu et al, Bioinformatics, 2009



Lots of Room for Improvement in Complex Prediction

Liu et al, Bioinformatics, 2009

scoring method: AdjustCD					*match_thres*=0.50							
					Aloy (#complexes: 63)				MIPS (#complexes: 162)			
clustering methods	k	#clusters	avg size	loc. score	#matched clusters	precision	#matched complexes	recall	#matched clusters	prec	#matched complexes	recall
CMC	0	172	9.83	0.823	53	0.308	53	0.841	42	0.244	55	0.340
	1	121	9.42	0.897	50	**0.413**	49	0.778	41	**0.339**	51	0.315
	2	148	8.50	0.899	57	0.385	**56***	**0.889**	44	0.297	**56***	**0.346**
	20	146	8.78	0.891	56	0.384	**56***	**0.889**	43	0.295	**56***	**0.346**
CFinder	0	103	13.84	0.528	39	0.379	38	0.603	34	0.330	40	0.247
	1	76	12.86	0.724	38	**0.500**	38	0.603	30	**0.395**	34	0.210
	2	95	11.66	0.713	44	0.463	**43**	**0.683**	36	0.379	46	0.284
	20	95	11.77	0.718	44	0.463	**43**	**0.683**	37	0.389	**49**	**0.302**
MCL	0	372	9.40	0.638	27	0.073	27	0.429	30	0.081	37	0.228
	1	120	10.18	0.848	49	0.408	49	0.778	40	0.333	**51**	**0.315**
	2	116	10.31	0.856	52	**0.448**	**52**	**0.825**	41	**0.353**	**51**	**0.315**
	20	110	10.75	0.849	49	0.445	49	0.778	37	0.336	47	0.290
MCode	0	61	7.31	0.849	20	0.328	20	0.317	18	0.295	22	0.136
	1	103	7.42	0.913	35	0.340	**35**	**0.556**	30	0.291	**39**	**0.241**
	2	88	8.67	0.897	34	**0.386**	34	0.540	29	**0.330**	**39**	**0.241**
	20	82	10.28	0.838	29	0.354	29	0.460	23	0.280	32	0.198

Table 3. The impact of the iterative scoring method on the performance of four clustering methods. For CMC, MCL and CFinder, we retain only the top-6000 interactions, and no new interactions are added. For MCode, we retain all the interactions with non-zero score and add top-3000 new interactions with the highest score. The 2nd column is the number of iterations *k* of the iterative scoring method, and *k*=0 means the PPI network is unweighted. The 3rd column is the number of clusters generated, the 4th and 5th column is the average size and co-localization score of generated clusters.

Inferring Protein Function

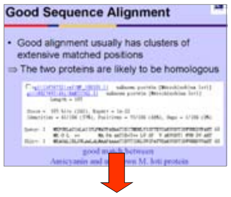
Function Assignment to Protein Seq

```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT
```

- **How do we attempt to assign a function to a new protein sequence?**

Guilt by Association of Seq Similarity

Compare *T* with seqs of known function in a db



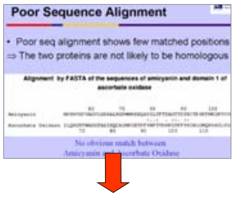
Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

[illegible]

good match between Amicyanin and unknown M. loti protein



Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

[illegible]

No obvious match between Amicyanin and Ascorbate Oxidase

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Discard this function as a candidate

What if there is no useful seq homolog?

- **Guilt by other types of association!**
  - Domain modeling (e.g., HMMPFAM)
  - Similarity of dissimilarities (e.g., SVM-PAIRWISE)
  - Similarity of phylogenetic profiles
  - Similarity of subcellular co-localization & other physico-chemico properties(e.g., PROTFUN)
  - Similarity of gene expression profiles
  - Similarity of protein-protein interaction partners
  - …
  - Fusion of multiple types of info

Similarity of Dissimilarities

Differences of “unknown” to other fruits are same as “apple” to other fruits

“unknown” is an “apple”!

	Orange$_1$	Banana$_1$	…

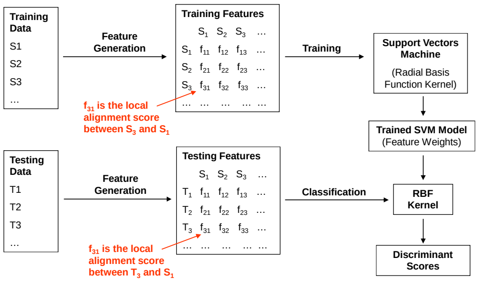
Apple$_1$	Color = red vs orange
	Skin = smooth vs rough
	Size = small vs small
	Shape = round vs round
		Color = red vs yellow
		Skin = smooth vs smooth
		Size = small vs small
		Shape = round vs oblong
			…

Orange$_2$	**Color = orange vs orange**
	Skin = rough vs rough
	Size = small vs small
	Shape = round vs round
		Color = orange vs yellow
		Skin = rough vs smooth
		Size = small vs small
		Shape = round vs oblong
			…

Unknown$_1$	**Color = red vs orange**
	Skin = smooth vs rough
	Size = small vs small
	Shape = round vs round
		Color = red vs yellow
		Skin = smooth vs smooth
		Size = small vs small
		Shape = round vs oblong
			…

…	…	…	…

SVM-Pairwise Framework



Training Data: S1, S2, S3, …

Feature Generation → Training Features

	S_1	S_2	S_3	…
S_1	f_{11}	f_{12}	f_{13}	…
S_2	f_{21}	f_{22}	f_{23}	…
S_3	f_{31}	f_{32}	f_{33}	…
…	…	…	…	…

f_{31} is the local alignment score between S_3 and S_1

Training → Support Vectors Machine (Radial Basis Function Kernel) → Trained SVM Model (Feature Weights)

Testing Data: T1, T2, T3, …

Feature Generation → Testing Features

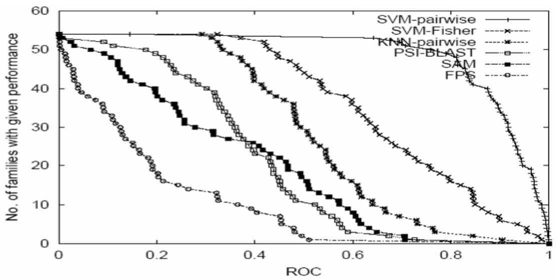
	S_1	S_2	S_3	…
T_1	f_{11}	f_{12}	f_{13}	…
T_2	f_{21}	f_{22}	f_{23}	…
T_3	f_{31}	f_{32}	f_{33}	…
…	…	…	…	…

f_{31} is the local alignment score between T_3 and S_1

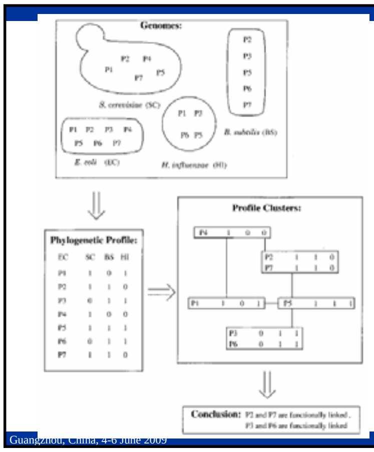
Classification → RBF Kernel → Discriminant Scores

Image credit: Kenny Chua

Performance of SVM-Pairwise



- ROC: The area under the curve derived from plotting true positives as a function of false positives for various thresholds



Phylogenetic Profile:

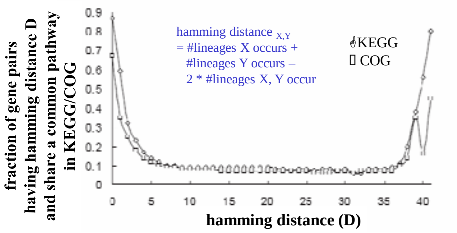
	EC	SC	BS	HI
P1		1	0	1
P2		1	1	0
P3		0	1	1
P4		1	0	0
P5		1	1	1
P6		0	1	1
P7		1	1	0

Conclusion: P2 and P7 are functionally linked, P3 and P6 are functionally linked

Phylogenetic Profiling: How It Works

Phylogenetic Profiling: Evidence

Wu et al., *Bioinformatics*, 19:1524--1530, 2003

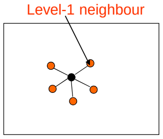


hamming distance $_{X,Y}$
= #lineages X occurs +
#lineages Y occurs –
2 * #lineages X, Y occur

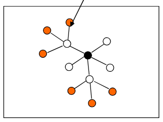
Functional Association Thru Interactions

- **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

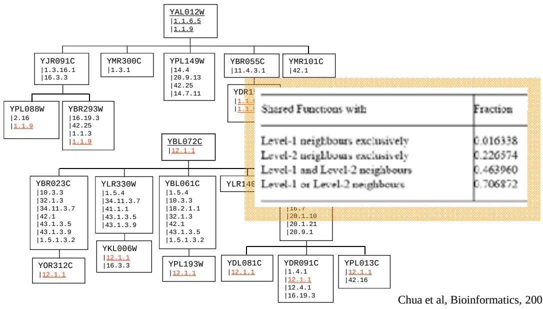
Level-1 neighbour



Level-2 neighbour



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

Chua et al, Bioinformatics, 2007

Functional Similarity Estimate: FS-Weighted Measure

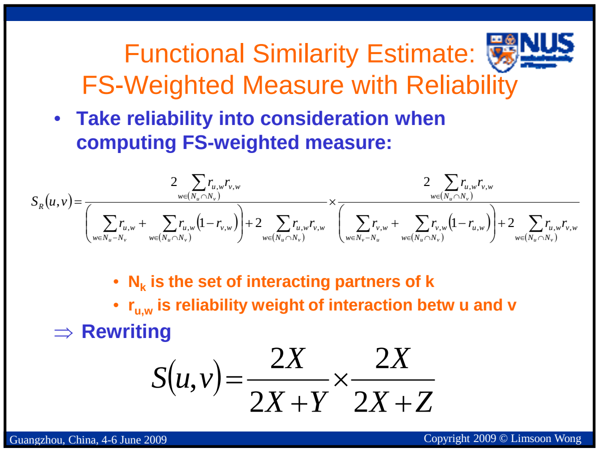
- **FS-weighted measure**

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

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

⇒**Rewriting this as**

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



Functional Similarity Estimate: FS-Weighted Measure with Reliability

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

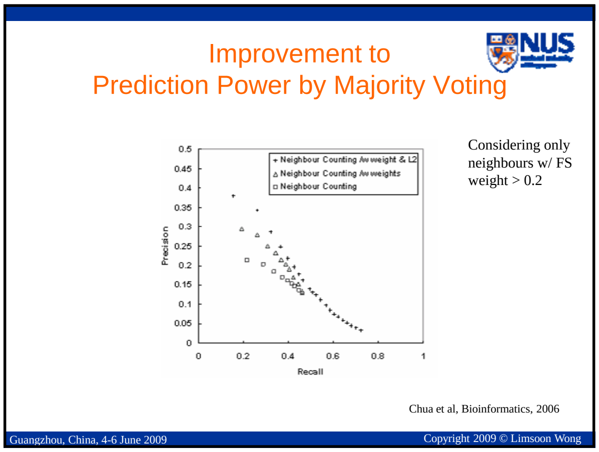
$$S_R(u,v)=\frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_u-N_v} r_{u,w}+\sum_{w\in(N_u\cap N_v)} r_{u,w}(1-r_{v,w})\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}\times\frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_v-N_u} r_{v,w}+\sum_{w\in(N_u\cap N_v)} r_{v,w}(1-r_{u,w})\right)+2\sum_{w\in(N_u\cap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

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

⇒ Rewriting

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

Guangzhou, China, 4-6 June 2009 Copyright 2009 © Limsoon Wong

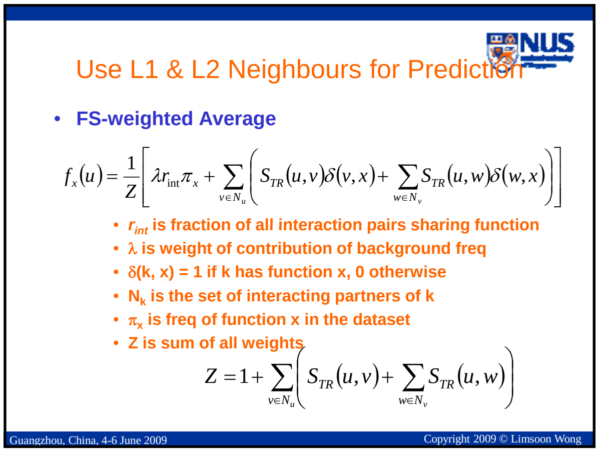


Improvement to Prediction Power by Majority Voting

Considering only neighbours w/ FS weight > 0.2

Chua et al, Bioinformatics, 2006

Guangzhou, China, 4-6 June 2009 Copyright 2009 © Limsoon Wong



Use L1 & L2 Neighbours for Prediction

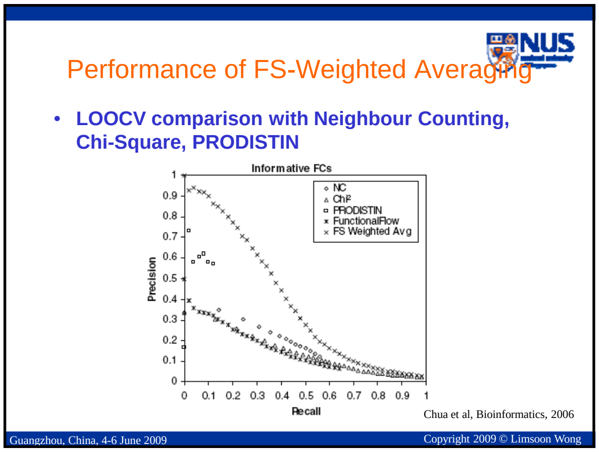
- **FS-weighted Average**

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

  - **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**

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

Guangzhou, China, 4-6 June 2009 Copyright 2009 © Limsoon Wong

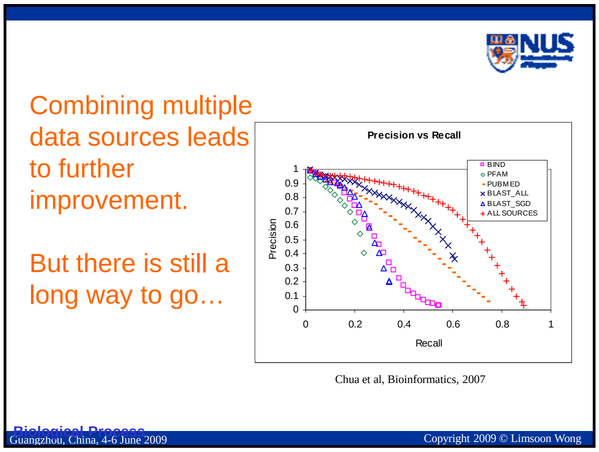


Performance of FS-Weighted Averaging

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

Chua et al, Bioinformatics, 2006

Guangzhou, China, 4-6 June 2009 Copyright 2009 © Limsoon Wong



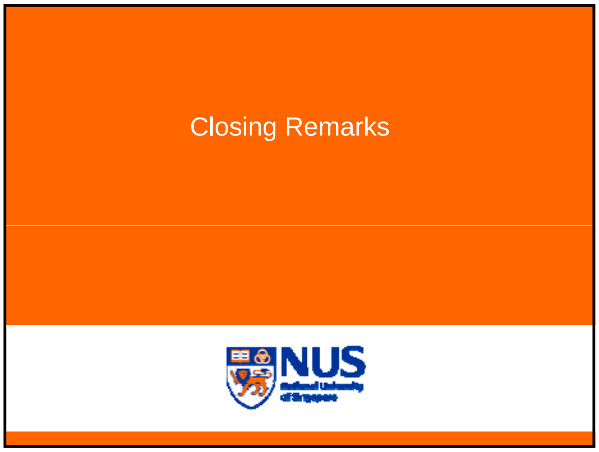
Combining multiple data sources leads to further improvement.

But there is still a long way to go…

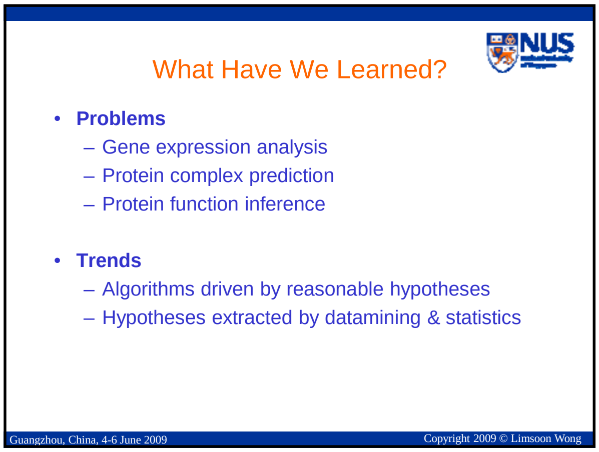
Chua et al, Bioinformatics, 2007

Biological Process

Guangzhou, China, 4-6 June 2009 Copyright 2009 © Limsoon Wong

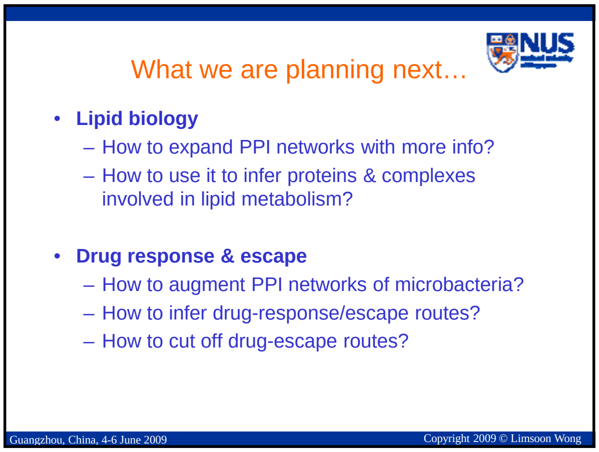


Closing Remarks



What Have We Learned?

- **Problems**
  - Gene expression analysis
  - Protein complex prediction
  - Protein function inference
- **Trends**
  - Algorithms driven by reasonable hypotheses
  - Hypotheses extracted by datamining & statistics



What we are planning next…

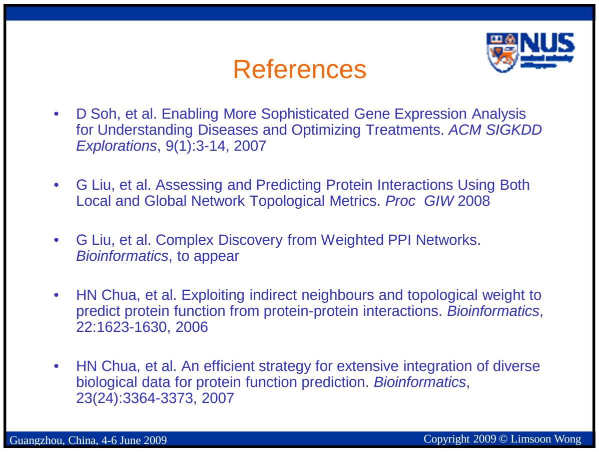
- **Lipid biology**
  - How to expand PPI networks with more info?
  - How to use it to infer proteins & complexes involved in lipid metabolism?
- **Drug response & escape**
  - How to augment PPI networks of microbacteria?
  - How to infer drug-response/escape routes?
  - How to cut off drug-escape routes?

Acknowledgements

- **Jinyan Li**
- **Huiqing Liu**

- **Difeng Dong**
- **Donny Soh**

- **Hon Nian Chua**
- **Guimei Liu**



References

- D Soh, et al. Enabling More Sophisticated Gene Expression Analysis for Understanding Diseases and Optimizing Treatments. *ACM SIGKDD Explorations*, 9(1):3-14, 2007
- G Liu, et al. Assessing and Predicting Protein Interactions Using Both Local and Global Network Topological Metrics. *Proc GIW* 2008
- G Liu, et al. Complex Discovery from Weighted PPI Networks. *Bioinformatics*, to appear
- HN Chua, et al. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22:1623-1630, 2006
- HN Chua, et al. An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, 23(24):3364-3373, 2007