Living with noise

Limsoon Wong

SinFra 2012

Paris, 15-16 Oct 2012

NUS
National University
of Singapore

Biology is full of noise

- **Experimental noise**
- **Intrinsic noise**

Living with noise

- When determining whether the value of a biological entity is above or below a threshold, instead of first determining its exact value and comparing that to the threshold, determine a distribution of that value and see whether it is likely to be above or below the threshold

- Instead of identifying and eliminating noise from samples, use bootstrap re-sampling to produce many bags of samples that are enriched with less noisy samples

- Use noise-robust logic reasoning

Batch Effect in Gene Expression Profiles

NUS
National University
of Singapore

Headaches in gene expression analysis

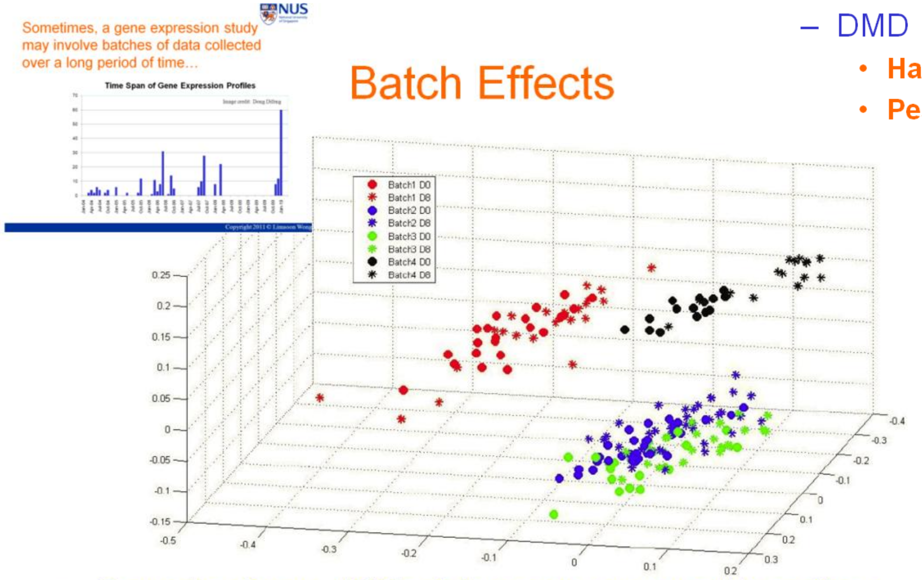
Percentage of Overlapping Genes

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



Sometimes, a gene expression study may involve batches of data collected over a long period of time…

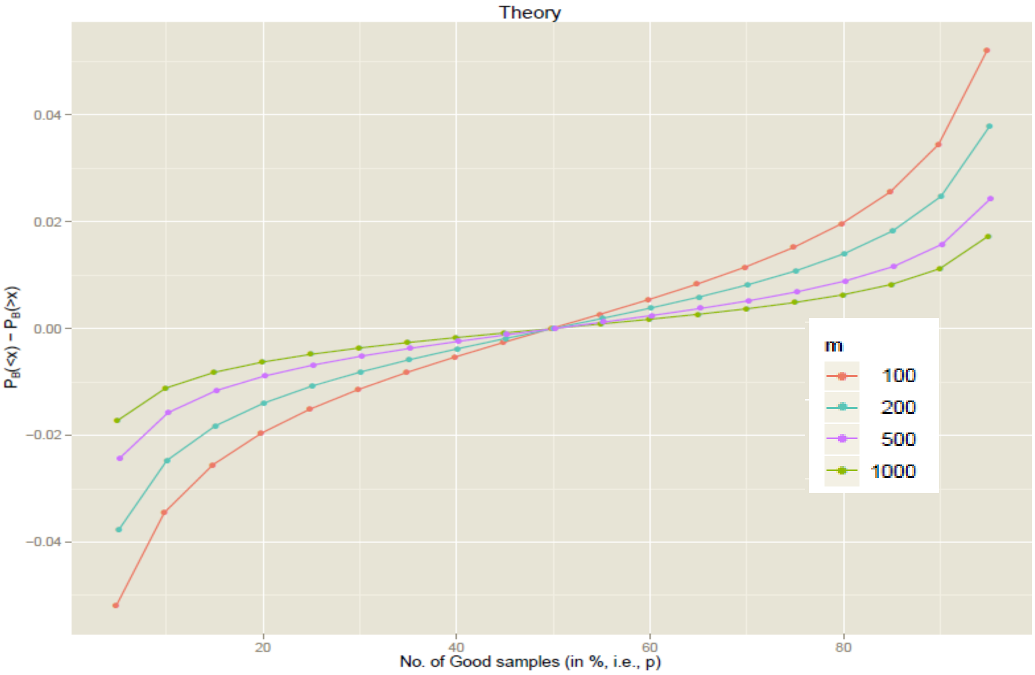
Batch Effects

- **Samples from diff batches are grouped together, regardless of subtypes and treatment response**

Image credit: Difeng Dong's PhD dissertation, 2011

Bootstrap sampling suppresses noise

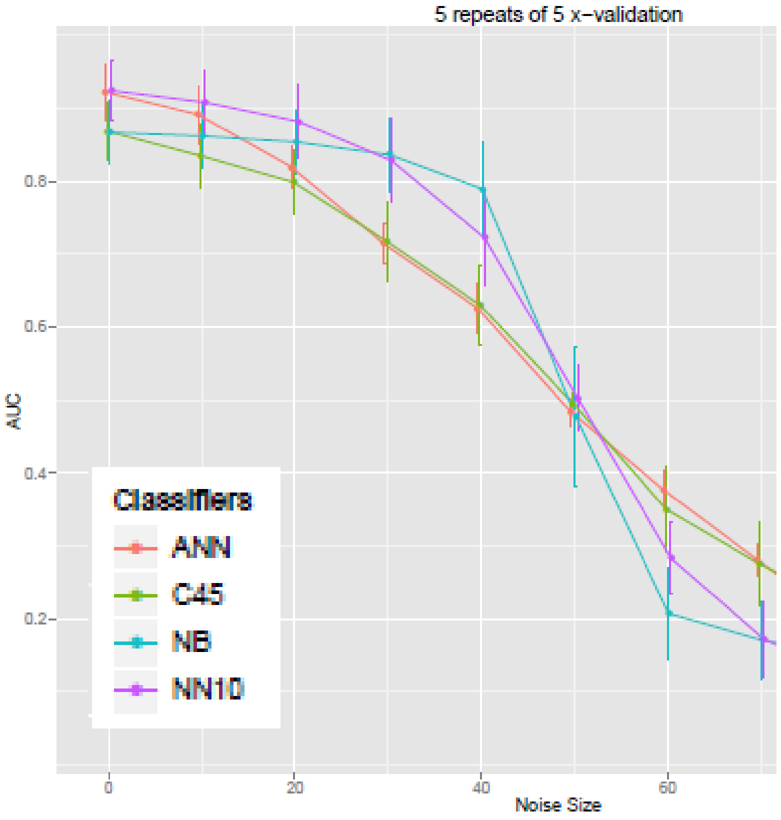
- **Suppose there are more “good” than “bad” samples in the training set**
- **Then any collection of its bootstrap replicates is likely to be enriched with bags containing more “good” than “bad” samples**



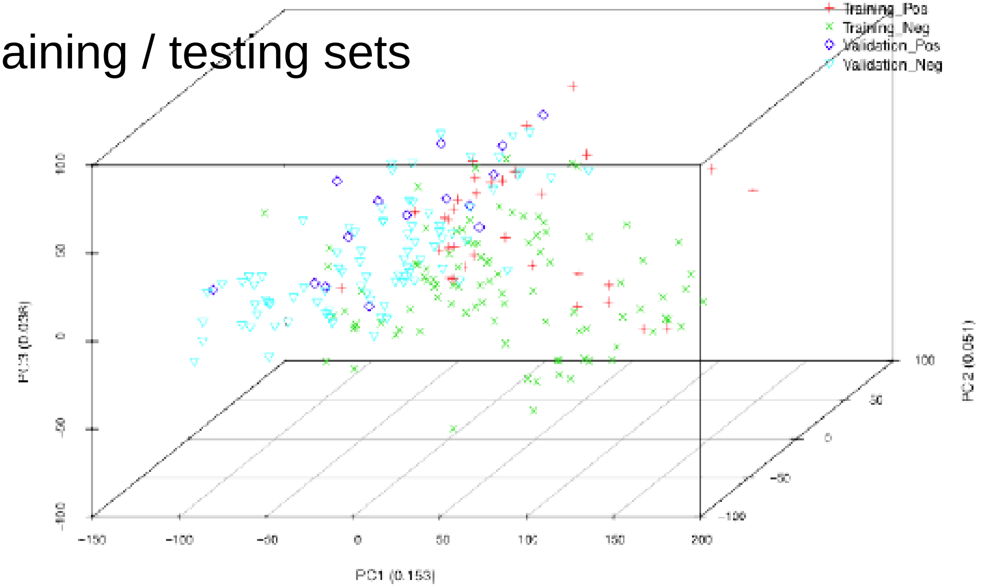
Why bagging works

- **Learning algo's are well behaved**

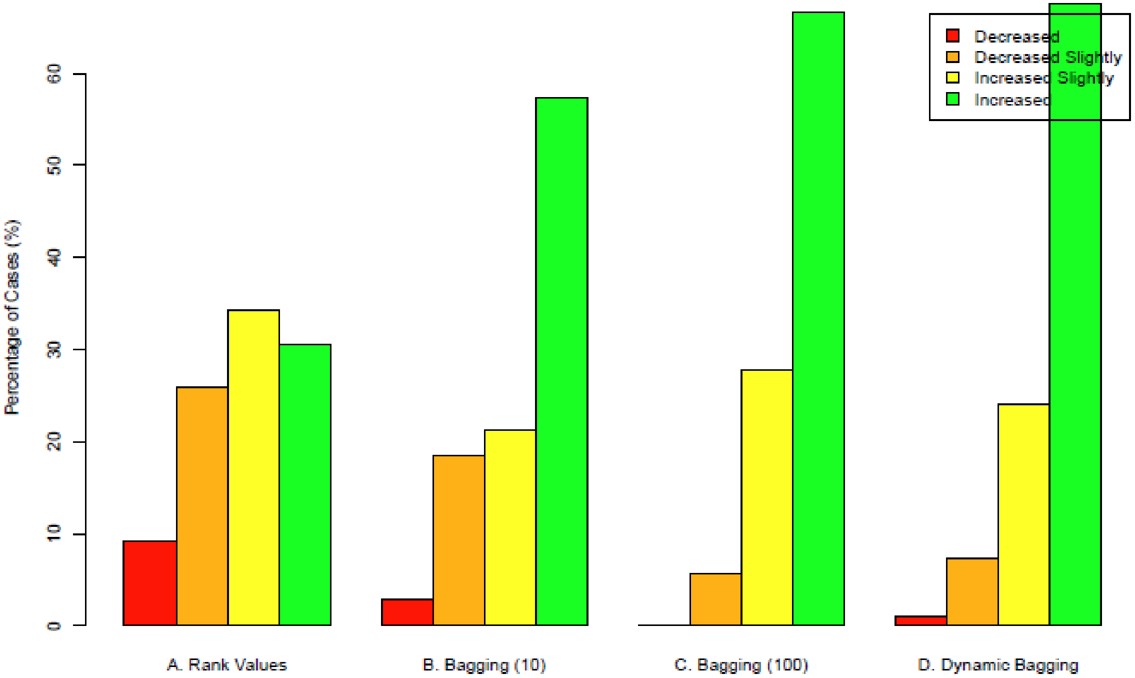
- **Given learning algo C and training set S with more "good" than "bad" samples. Let B_1, ..., B_n be boostrap replicates of S. Then a bagging classifier based on a majority vote of classifiers $C(B_1)$, ..., $C(B_n)$ is better than C(S)**



Batch effect in training / testing sets



Significantly improves cross-batch prediction accuracy in gene expression profile analyses



Protein Interactome Cleansing

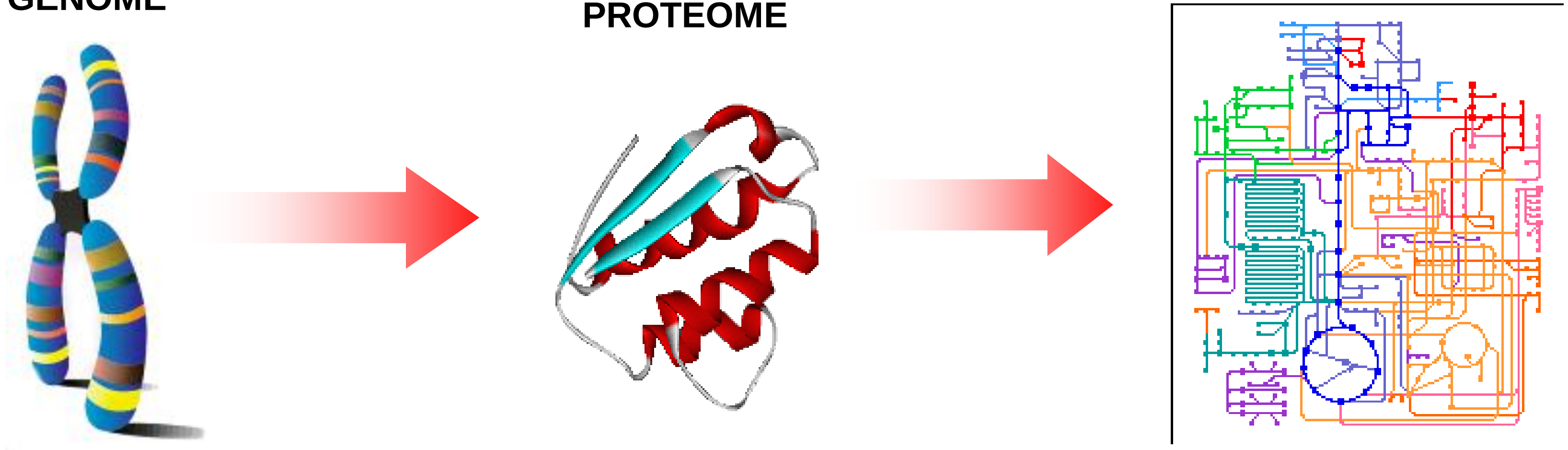
National University of Singapore

Why Biological Networks?

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

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

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



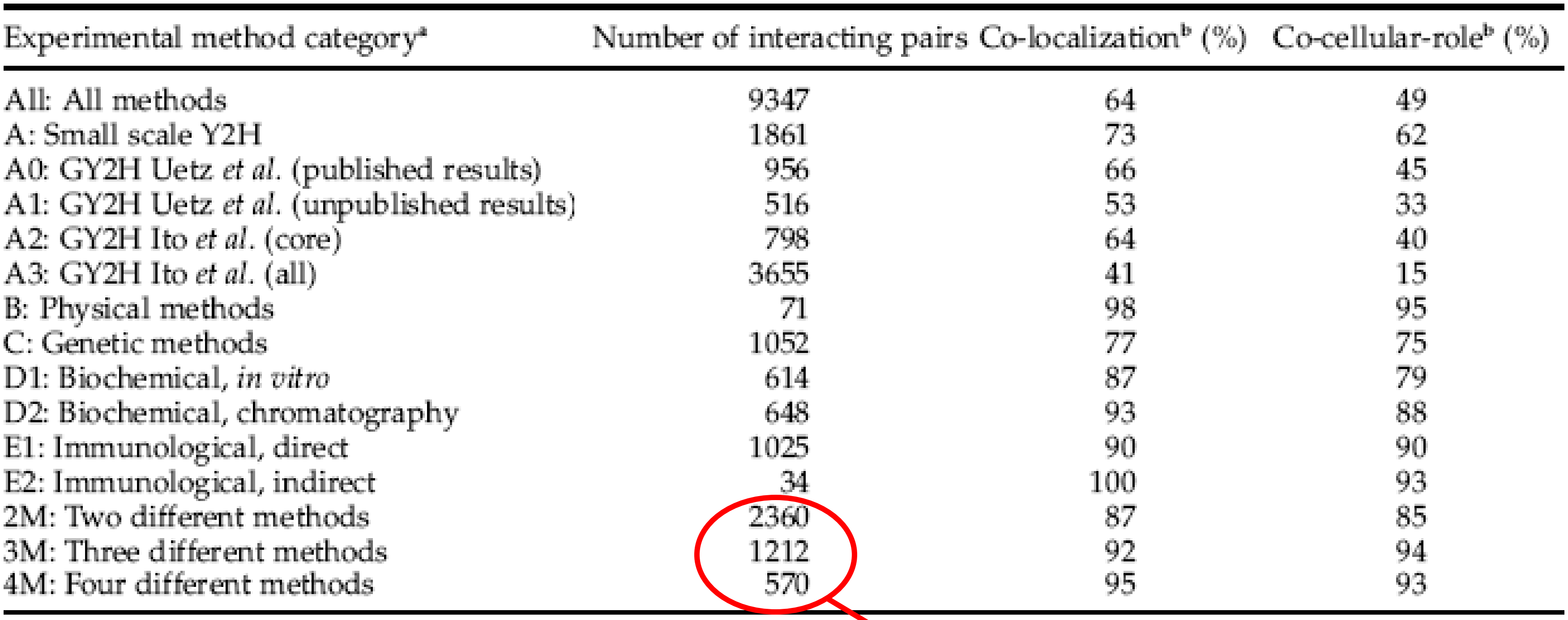
GENOME

PROTEOME

“INTERACTOME”

Slide credit: See-Kiong Ng

Identifying true PPIs in noisy expts



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

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

Large disagreement betw methods

- **PPIs are the basis of many biological mechanisms**
- **But there is a lot of noise in high-throughput PPI assays**

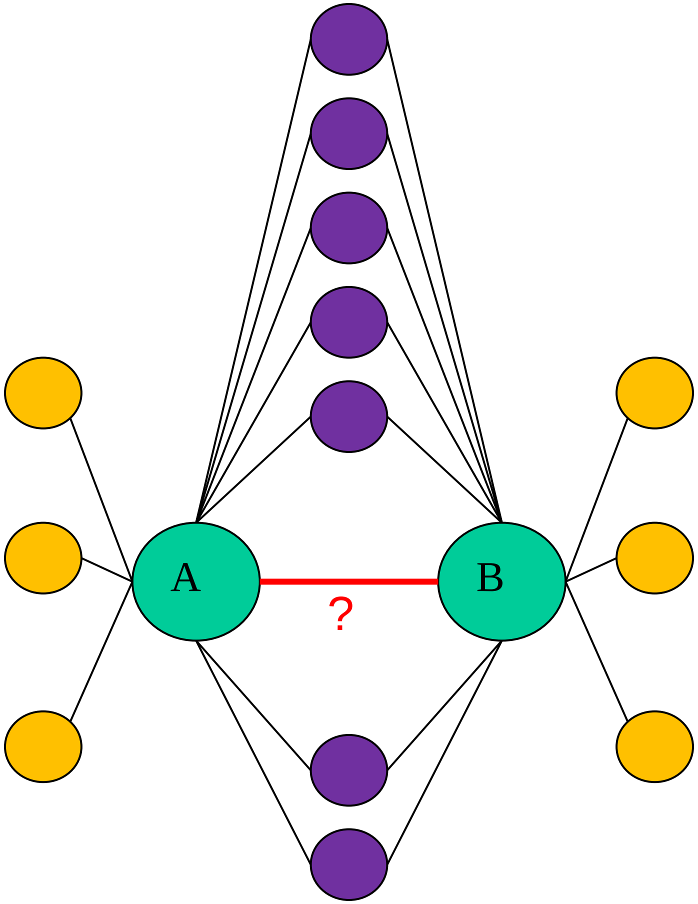
Can noise be removed w/o more info?

- **Some common ideas to remove noise**
  - A PPI detected by two independent assays is more likely to be true
  - Two proteins participating in same biological process are more likely to interact
  - Two proteins in the same cellular compartments are more likely to interact

- **But these need additional expt and additional info**

- **Can we do better?**

Topology of neighbourhood of real PPIs



- **Suppose 20% of putative PPIs are noise**

⇒ **≥ 3 purple proteins are real partners of both A and B**

⇒ **A and B are likely localized to the same cellular compartment (Why?)**

- **Fact: Proteins in the same cellular compartment are 10x more likely to interact than other proteins**

⇒ **A and B are likely to interact**

Liu et al. **Complex discovery from weighted PPI networks**. *Bioinformatics*, 25:1891-1897, 2009

Iterated CD Distance

- CD-distance

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

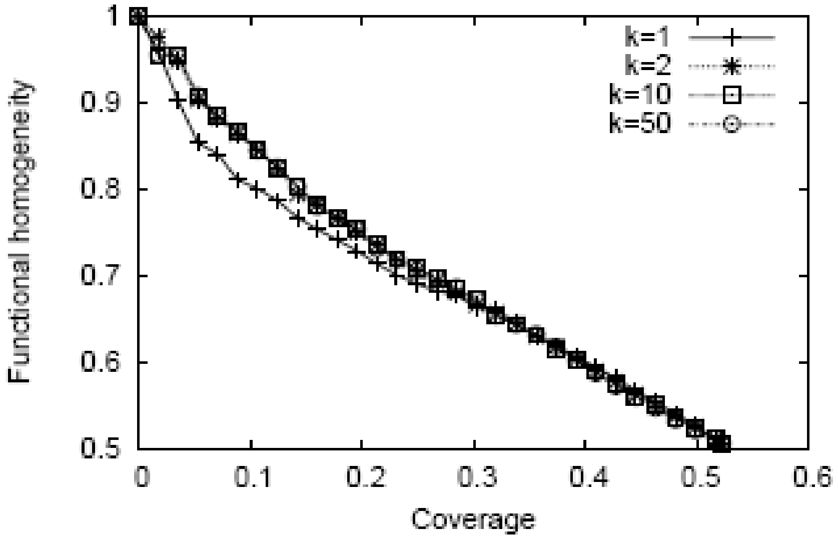
- X is # common neighbours of 1st & 2nd proteins
- Y/Z is # unique neighbours of 1st/2nd protein

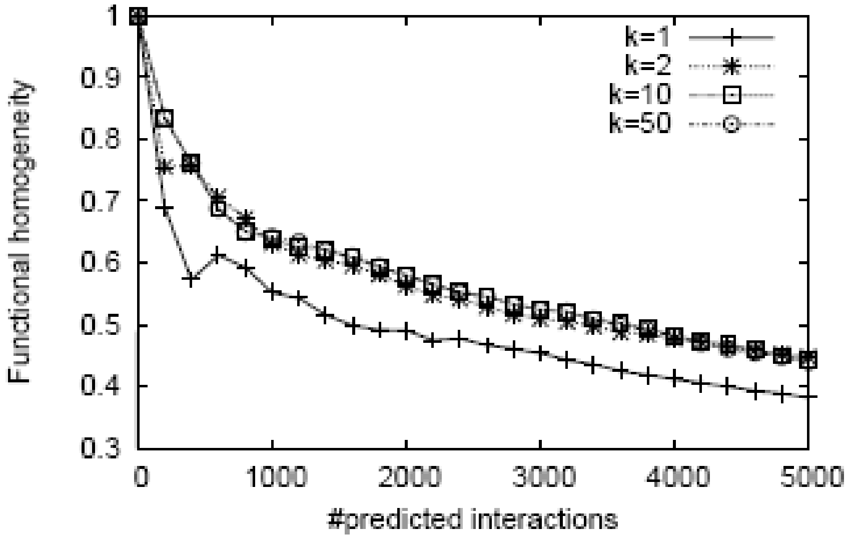
- These counts are noisy. ∴ Use CD-distance to weigh these counts and recompute CD-distance

$$w_L^k(u,v)=\frac{\sum_{x\in N_u\cap N_v} w_L^{k-1}(x,u)+\sum_{x\in N_u\cap N_v} w_L^{k-1}(x,v)}{\sum_{x\in N_u} w_L^{k-1}(x,u)+\sum_{x\in N_v} w_L^{k-1}(x,v)+\lambda_u^k+\lambda_v^k}$$

Performance wrt Functional Homogeneity

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





- **Ditto wrt localization coherence (not shown)**

Consistency of Proteomic Profiles

NUS

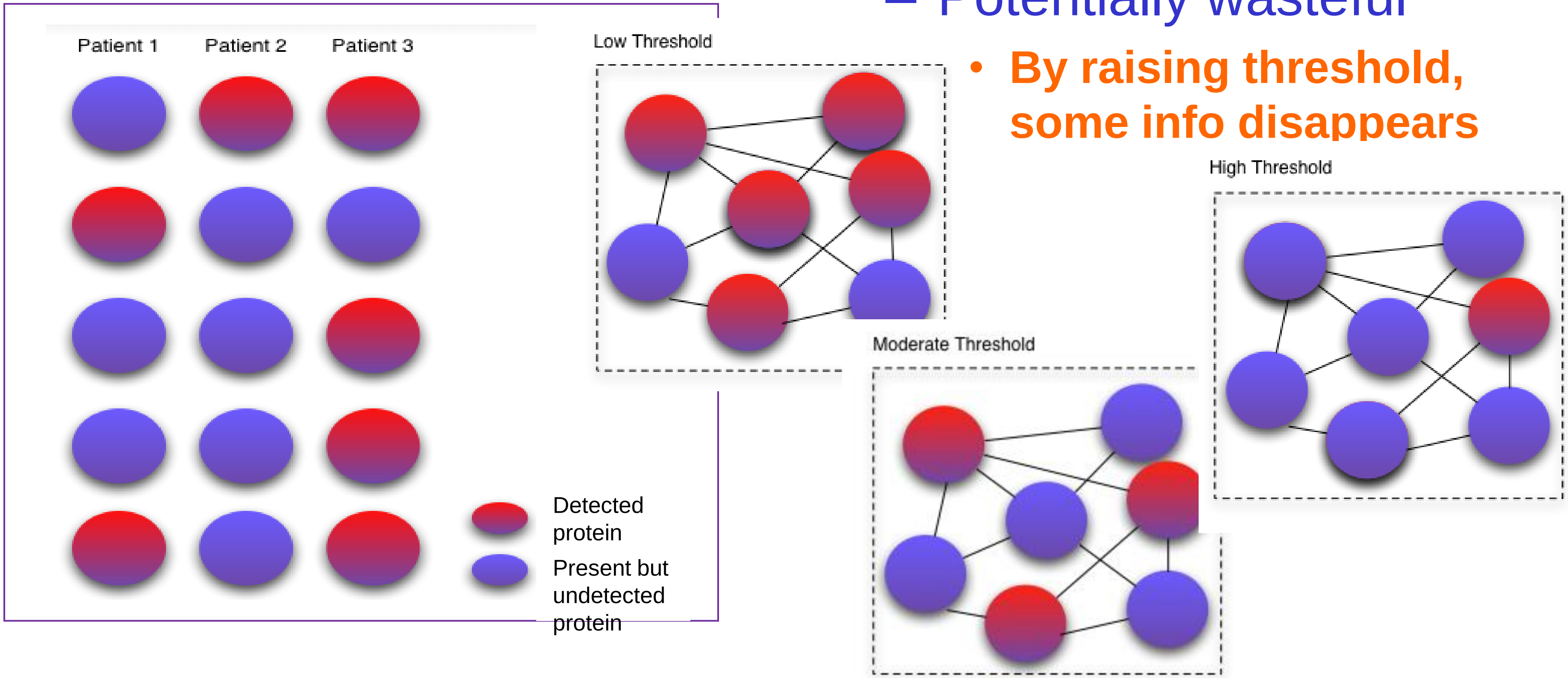
National University of Singapore

Issues in Proteomic Profiling

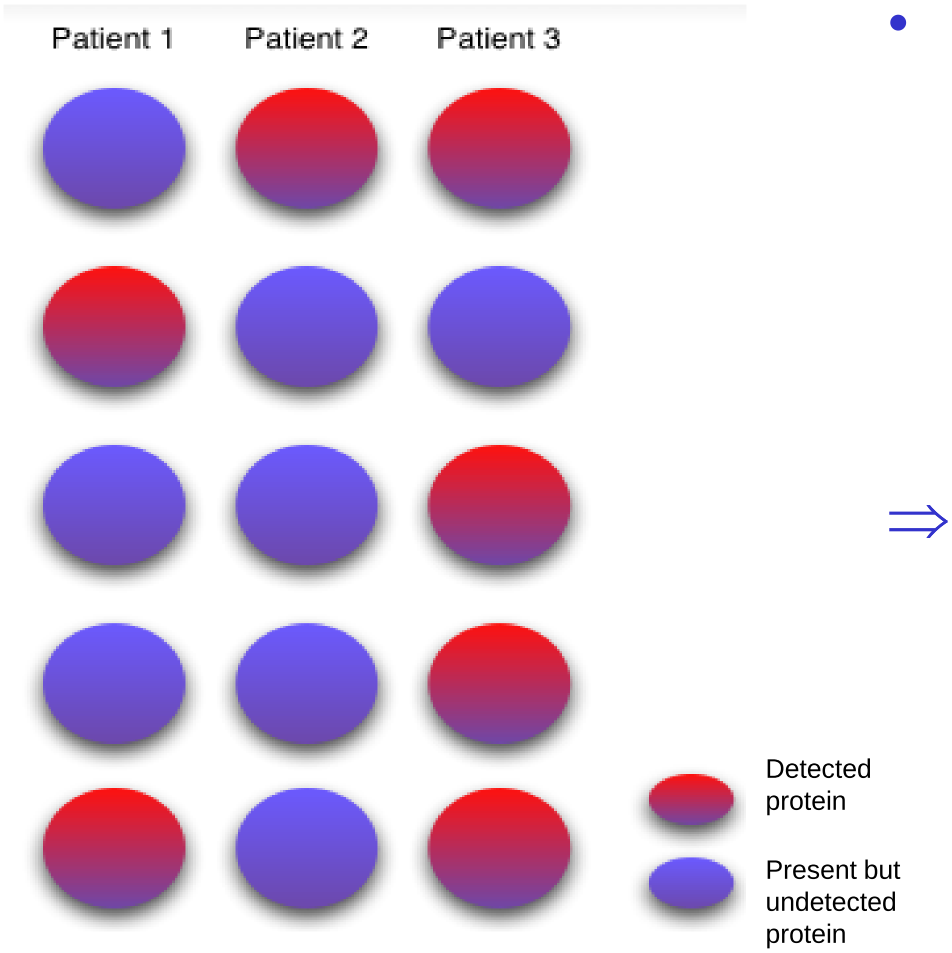
- **Coverage**
- **Consistency**

⇒Thresholding

- – Somewhat arbitrary
- – Potentially wasteful
  - **By raising threshold, some info disappears**



Intuitive Example



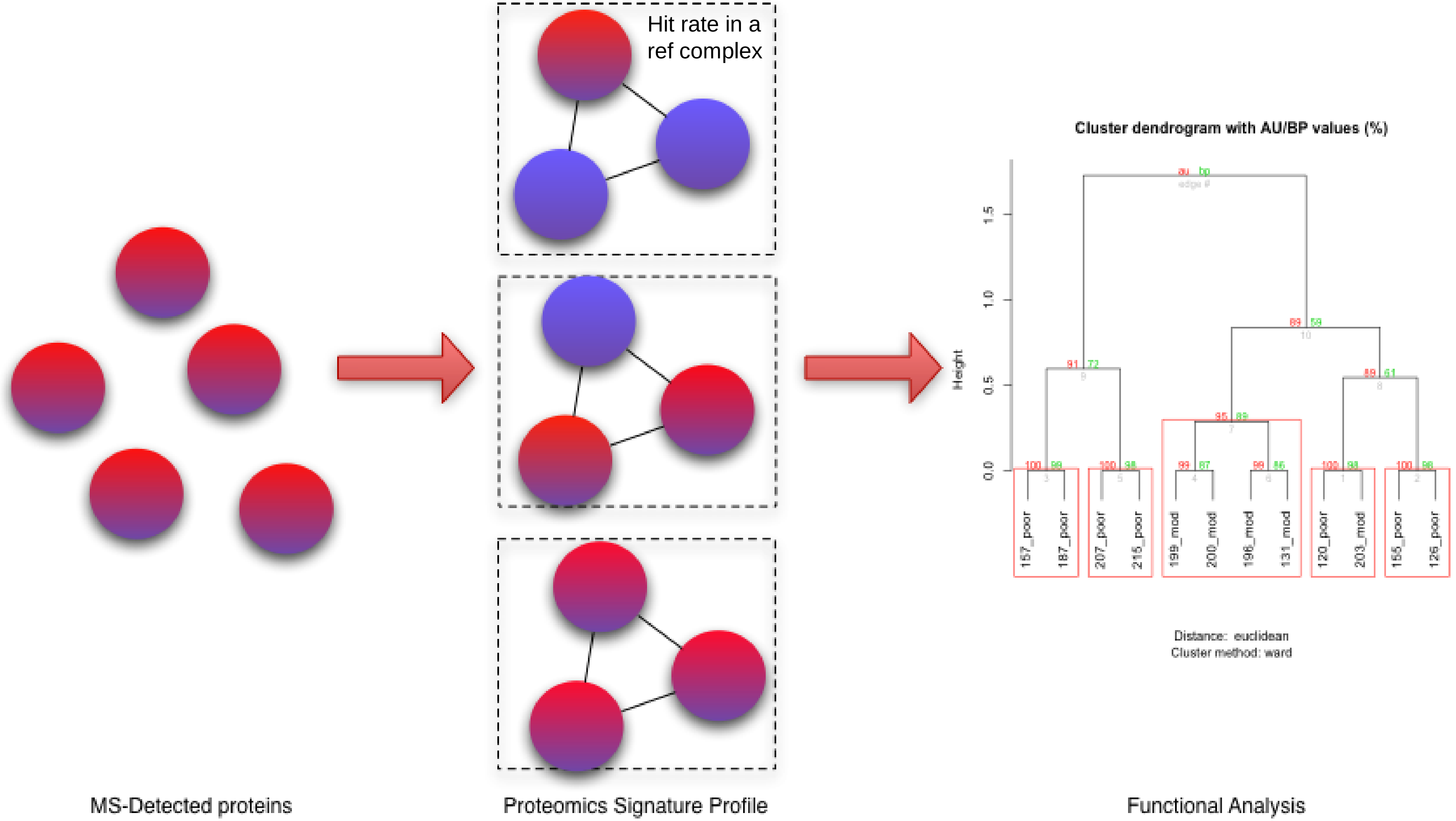
- **Suppose the failure to form a protein complex causes a disease**
  - If any component protein is missing, the complex can’t form

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

  - Construct a profile based on complexes?

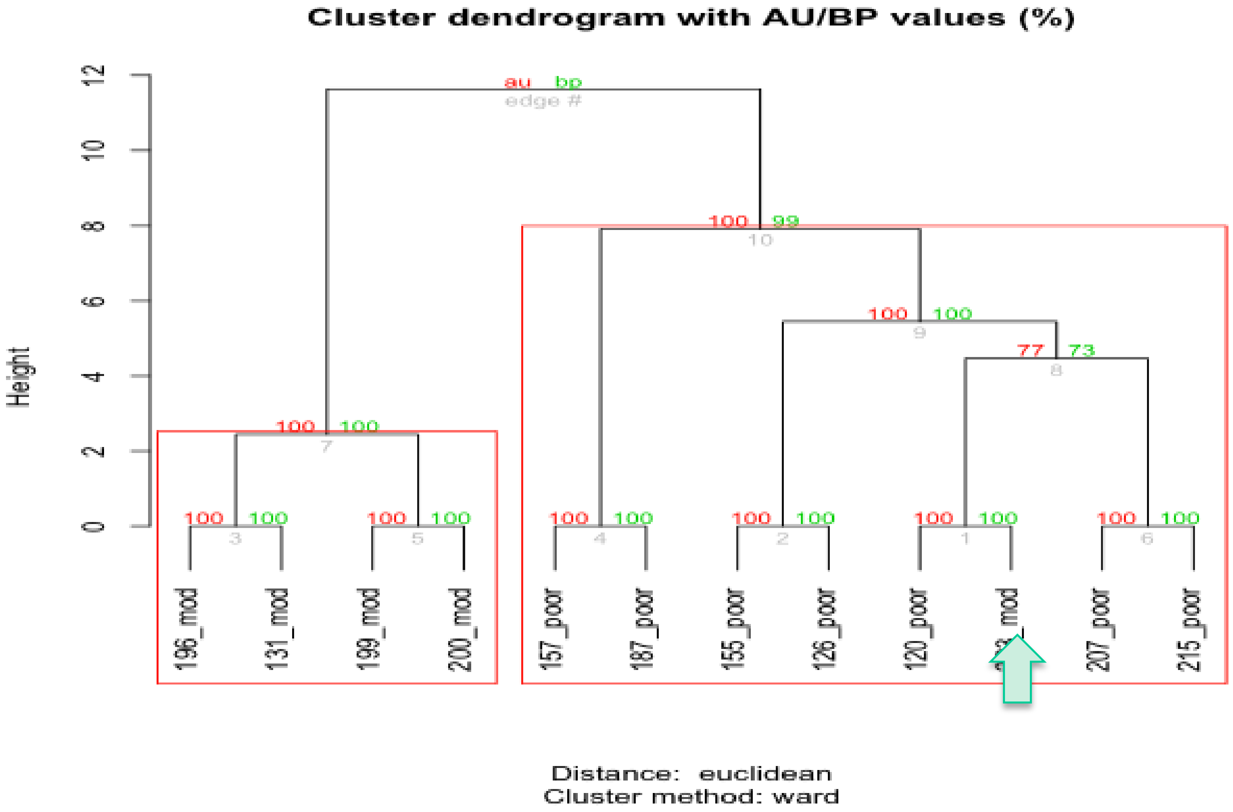
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. 11(3):1571-1581, 2012.

"Threshold-free" Principle of PSP



Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. 11(3):1571-1581, 2012.

Consistency: Samples segregate by their classes with high confidence



References & Acknowledgements

- **Materials for this talk are from joint works with my students (Kenny Chua, Wilson Goh, Chuan Hock Koh) and postdoc (Guimei Liu):**
  - Liu et al. “Complex discovery from weighted PPI networks”. *Bioinformatics*, 25:1891-1897, 2009
  - Goh et al. “Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics”. *Journal of Proteome Research*, 11(3):1571-1581, 2012
  - Koh & Wong. “Embracing noise to improve cross-batch prediction accuracy”. Manuscript, 2012