Computational Thinking in Genome and Proteome Analysis:

A Logician's Adventures in Computational Biology

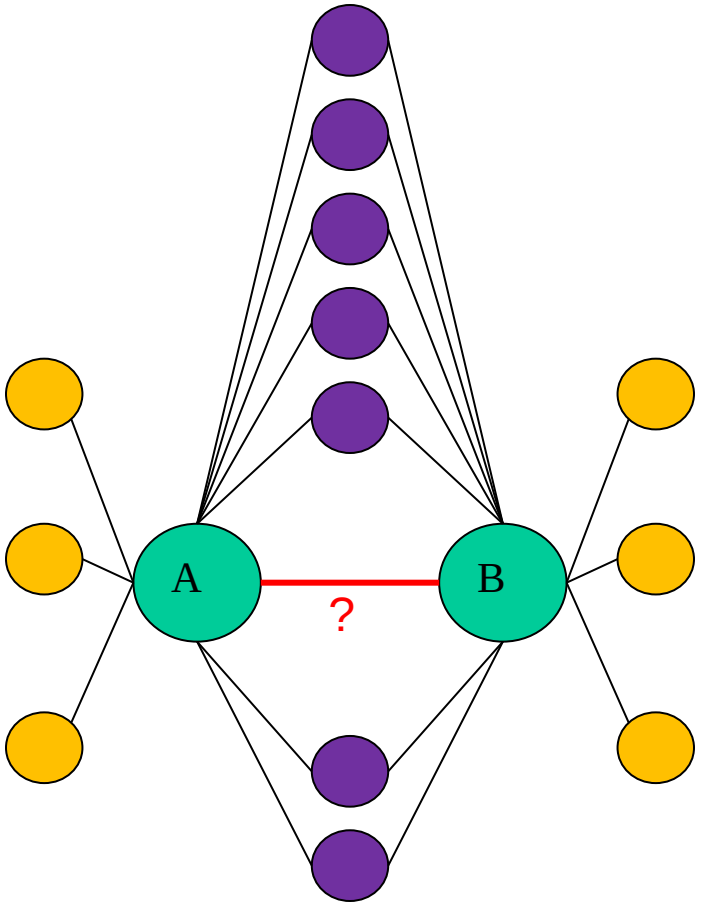
Wong Limsoon

NUS
National University
of Singapore

NUS National University of Singapore

what computational thinking is

An example of computational thinking



- **Suppose 20% noise in the PPIN**

⇒ **≥ 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**

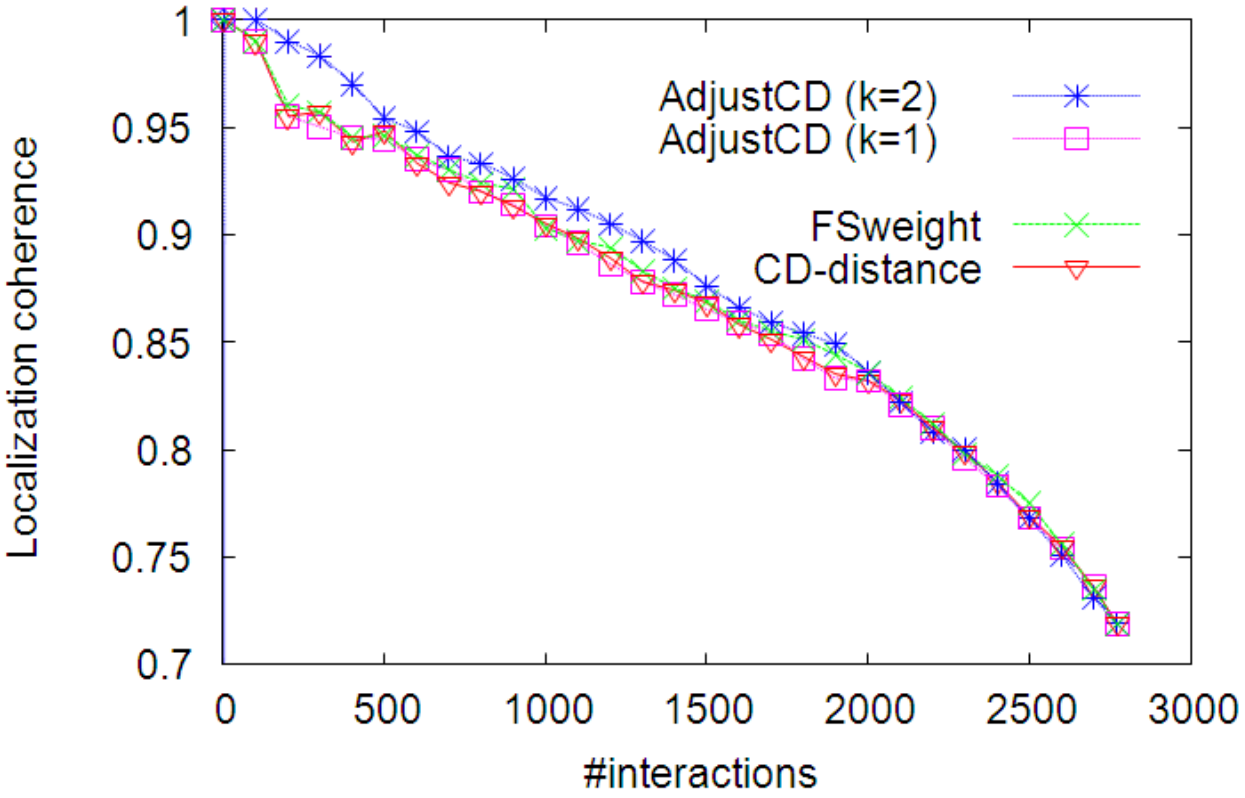
⇒ **A and B are likely to interact**

⇒ **CD distance measures the proportion of A and B's neighbours that are common between them**

Chua & Wong. Increasing the Reliability of Protein Interactomes.
Drug Discovery Today, 13(15/16):652--658, 2008

Successful noise reduction in PPIN

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



NUS National University of Singapore

gene expression profile analysis

Irreproducibility Issue

- **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

Individual Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **How many genes on a microarray are expected to perfectly correlate to these samples?**

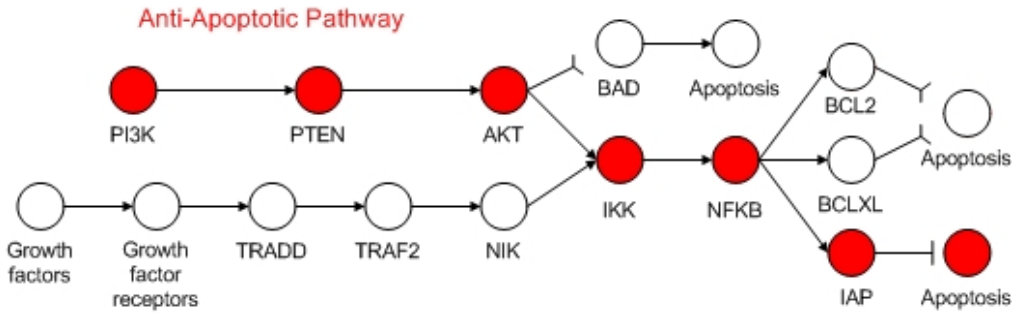
- **Prob(a gene is correlated) = $1/2^6$**
- **# of genes on array = 100,000**

⇒ **E(# of correlated genes) = 1,562**

⇒ **Many false positives**

- **These cannot be eliminated based on pure statistics!**

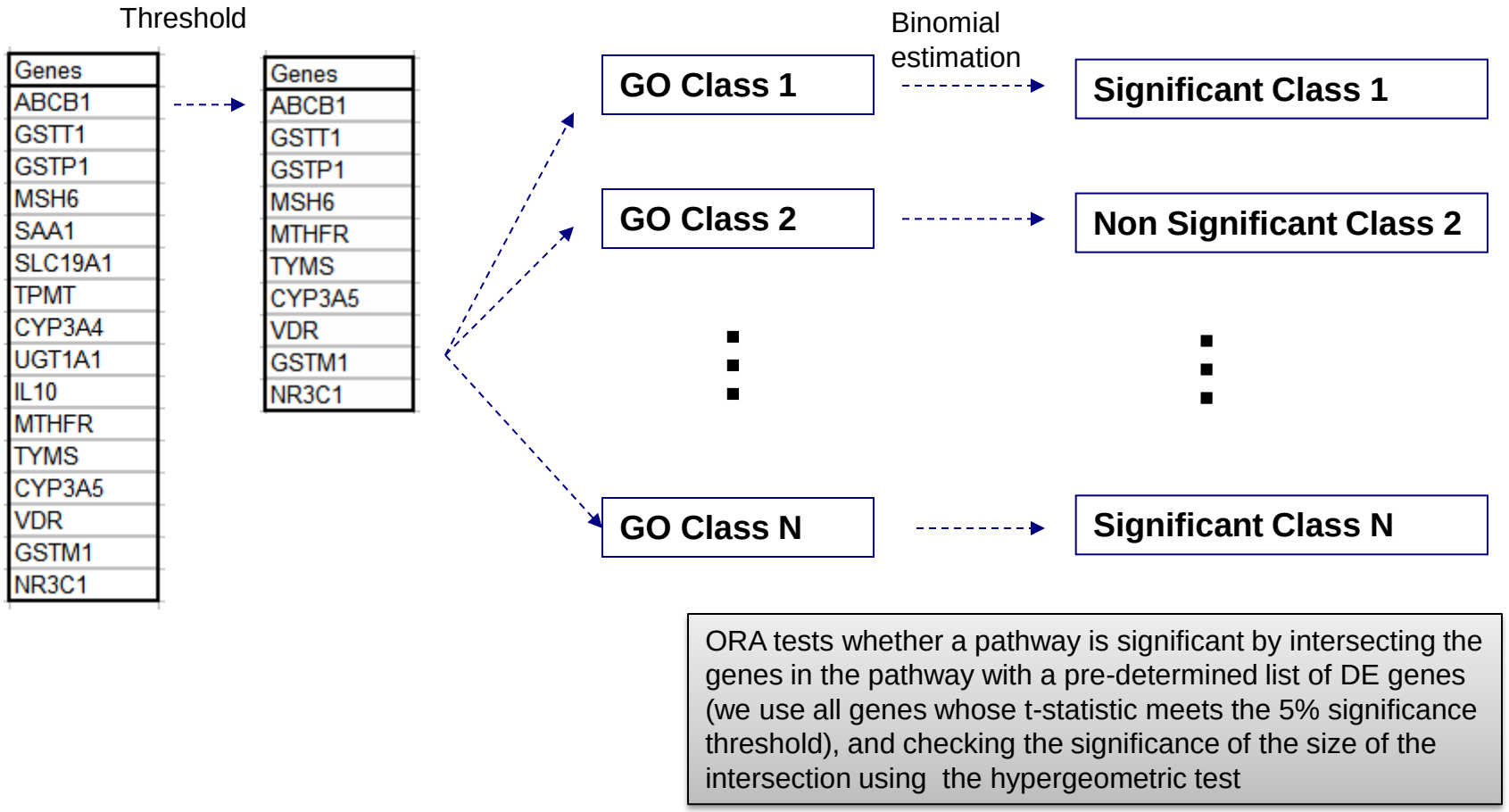
Gene Regulatory Circuits



- **Each disease phenotype has some underlying cause**
- **There is some 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**

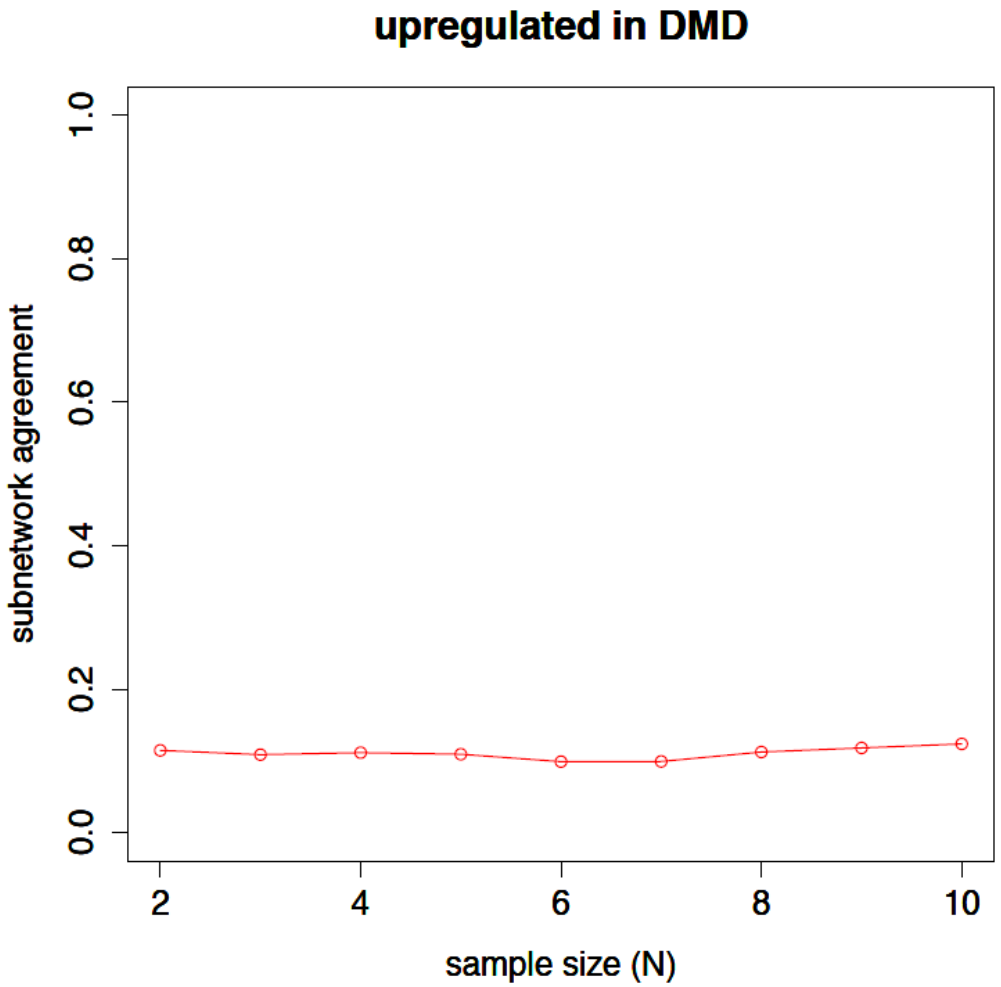
Overlap Analysis: ORA



ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (we use all genes whose t-statistic meets the 5% significance threshold), and checking the significance of the size of the intersection using the hypergeometric test

S Draghici et al. “Global functional profiling of gene expression”. *Genomics*, 81(2):98-104, 2003.

Disappointing Performance



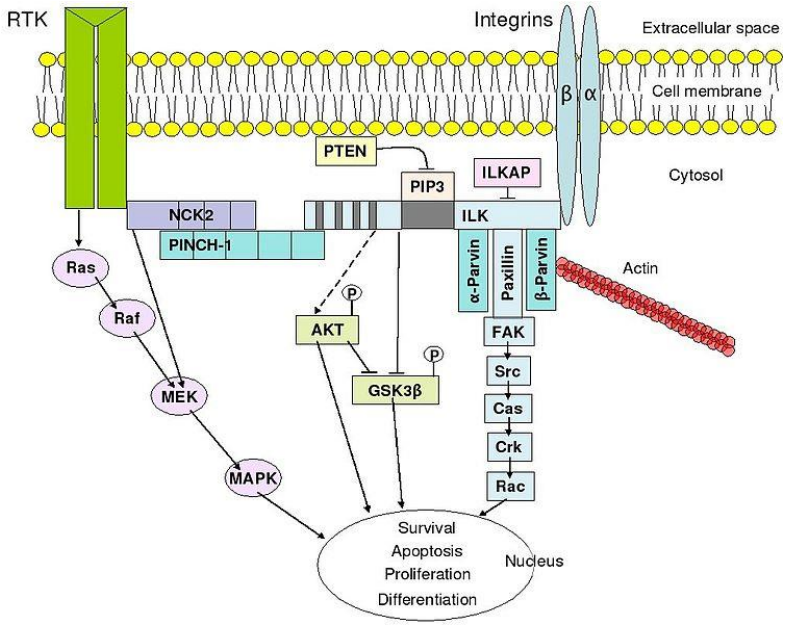
DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010

Issue #1 with ORA

- **Its null hypothesis basically says "Genes in the given pathway behaves no differently from randomly chosen gene sets of the same size"**

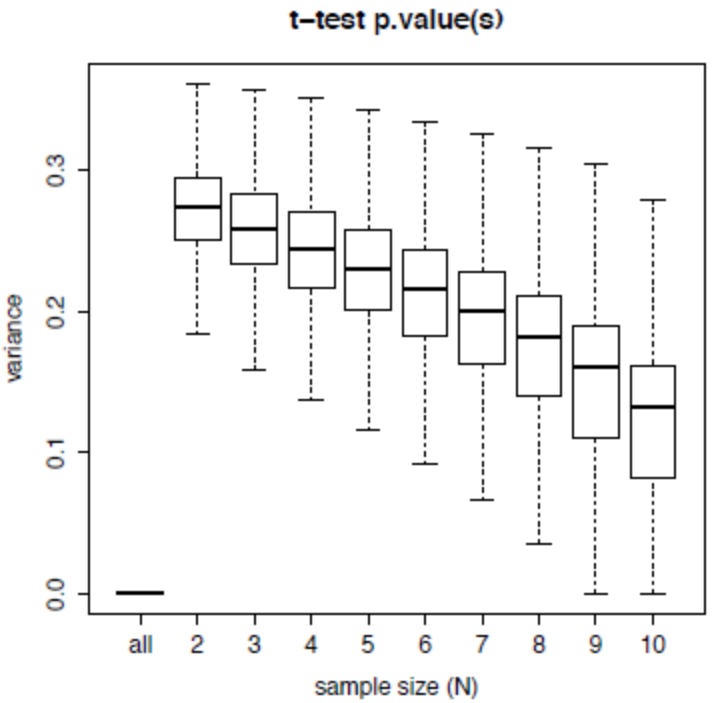
- **This may lead to lots of false positives**



- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behavour of genes in a pathway is more coordinated than random ones

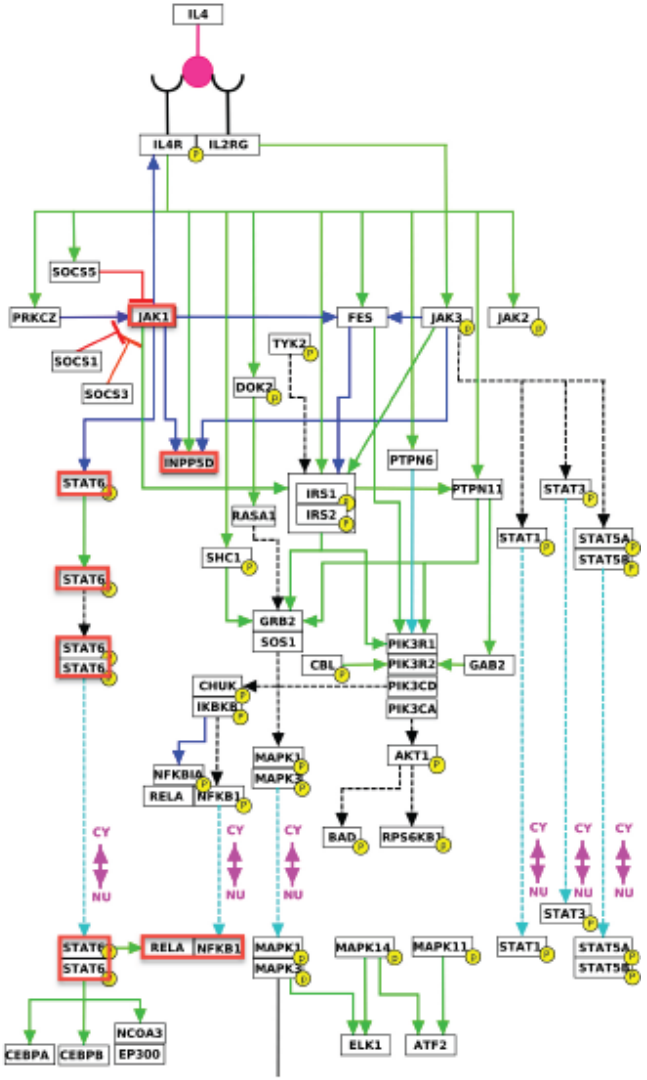
Issue #2 with ORA

- **It relies on a pre-determined list of DE genes**
- **This list is sensitive to the test statistic used and to the significance threshold used**
- **This list is unstable regardless of the threshold used when sample size is small**



Issue #3 with ORA

- **It tests whether the entire pathway is significantly differentially expressed**
- **If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch**

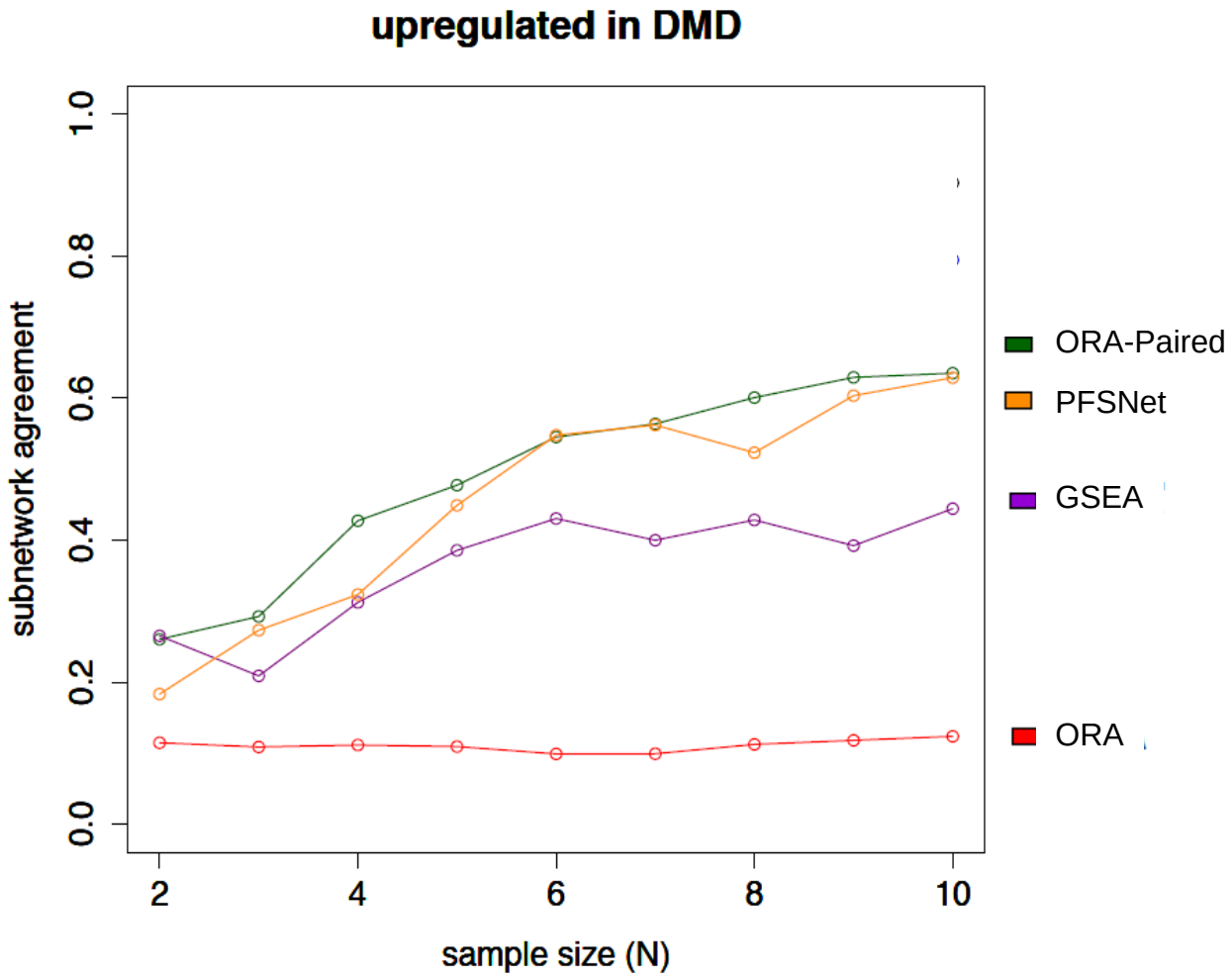


ORA-Paired: Paired Test and New Null Hypothesis

- Let g_i be genes in a given pathway P
- Let p_j be patients
- Let q_k be normals
- Let $\Delta_{i,j,k} = Expr(g_i,p_j) – Expr(g_i,q_k)$
- Test whether $\Delta_{i,j,k}$ is a ditribution with mean 0

- **Issue #1 is solved**
  - The null hypothesis is now “If a pathway P is irrelevant to the difference between patients and normals, then the genes in P are expected to behave similarly in patients and normals”
- **Issue #2 is solved**
  - No longer need a pre-determined list of DE genes
  - Sample size is now much larger
    - ~~# patients + # normals~~
    - # patients * # normals * # genes in P

Much Better Performance

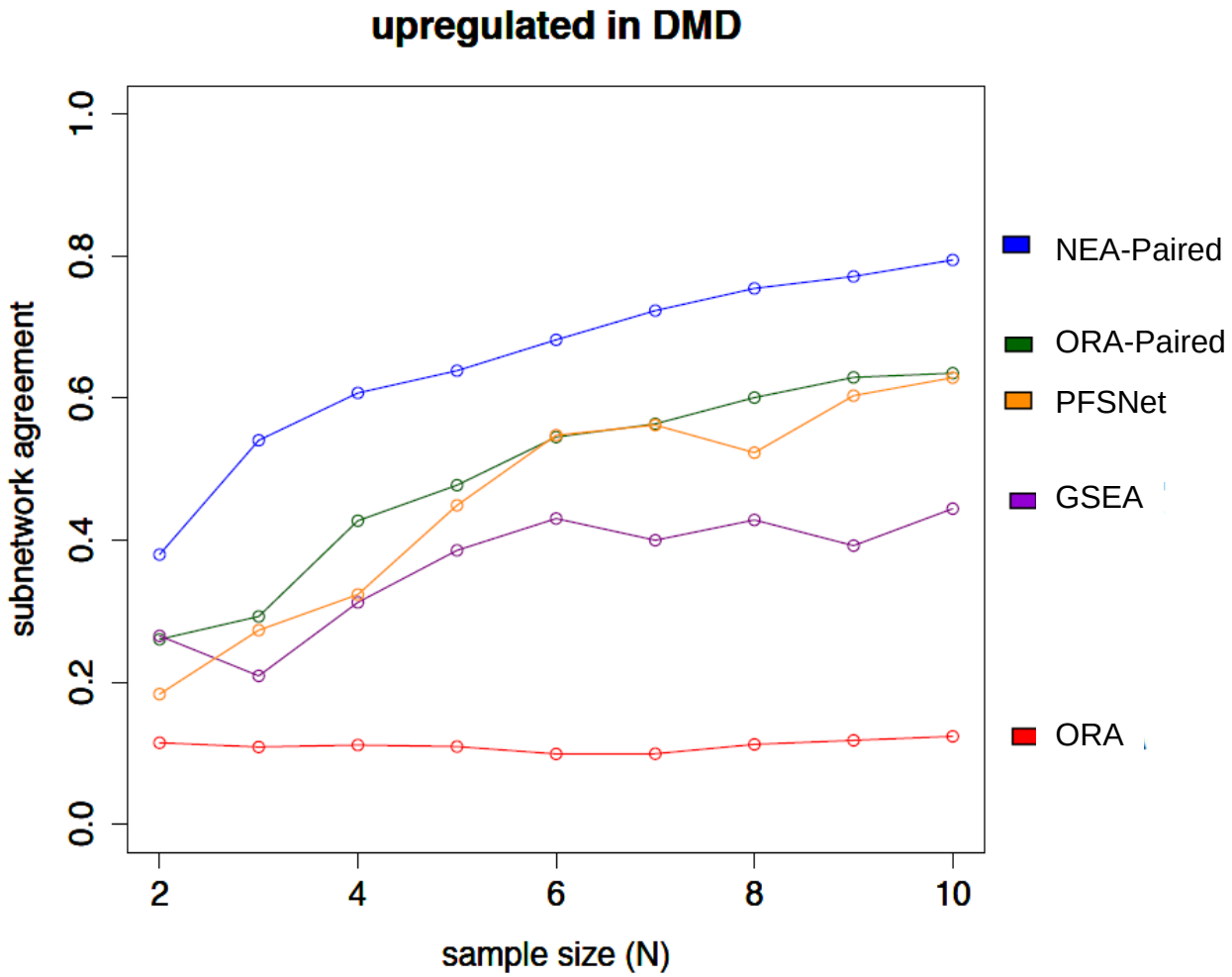


NEA-Paired: Paired Test on Subnetworks

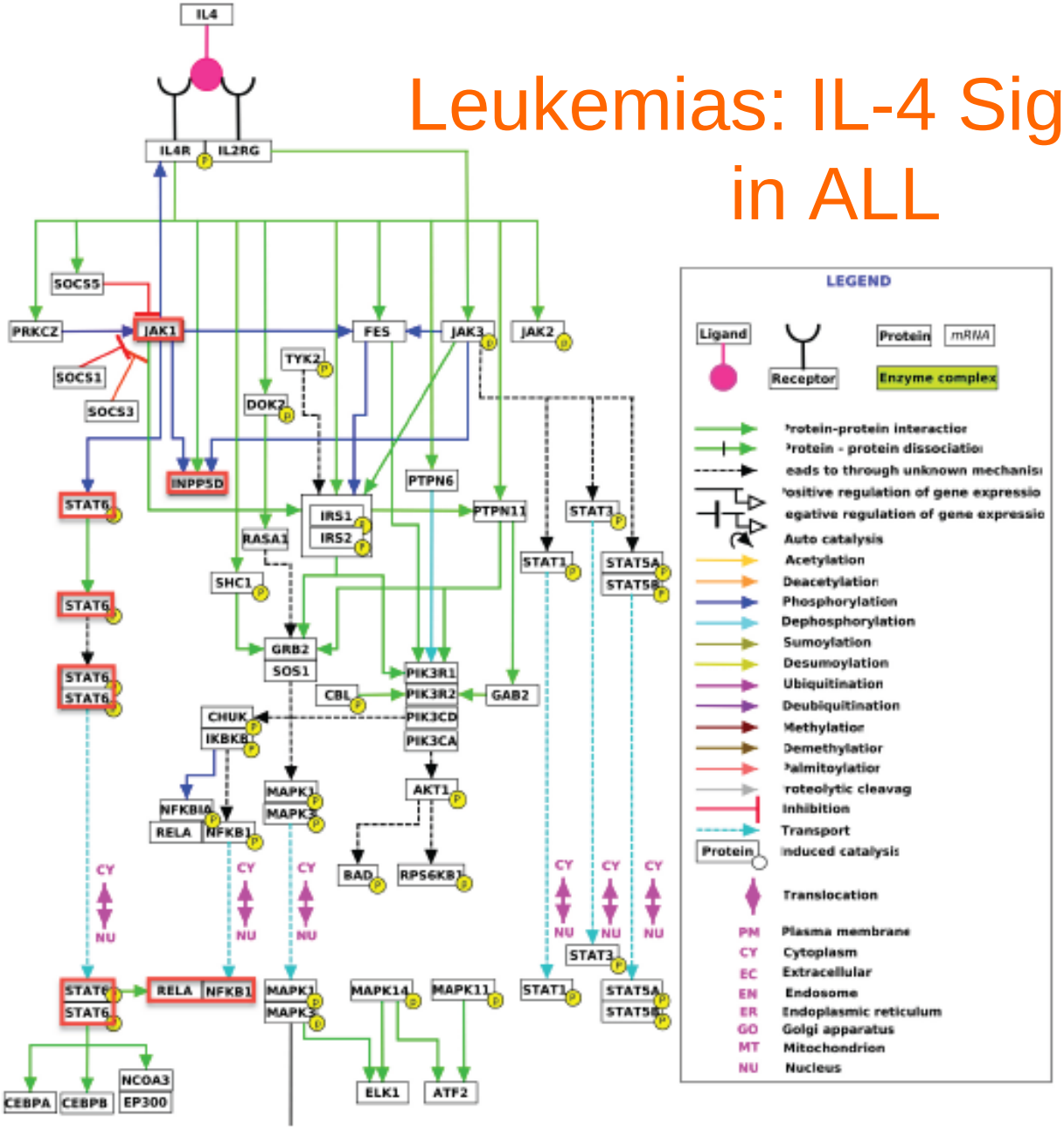
- **Given a pathway P**
- **Let each node and its immediate neighbourhood in P be a subnetwork**
- **Apply ORA-Paired on each subnetwork individually**

- **Issues #1 & #2 are solved as per ORA-Paired**
- **Issue #3 is partly solved**
  - Testing subnetworks instead of whole pathways
  - But subnetworks derived in fragmented way

Even Better Performance



Leukemias: IL-4 Signaling in ALL



For the Leukemia dataset (in which patients are either classified to have acute lymphoblastic leukemia or acute myeloid leukemia), one of the significant subnetworks that is biologically relevant is part of the Interleukin-4 signaling pathway; see figure 6b (supplementary material). The binding of Interleukin-4 to its receptor (Cardoso *et al.*, 2008) causes a cascade of protein activation involving JAK1 and STAT6 phosphorylation. STAT6 dimerizes upon activation and is transported to the nucleus and interacts with the RELA/NFKB1 transcription factors, known to promote the proliferation of T-cells (Rayet and Gelinas, 1999). In contrast, acute myeloid leukemia does not have genes in this subnetwork up-regulated and are known to be unrelated to lymphocytes.

NUS National University of Singapore

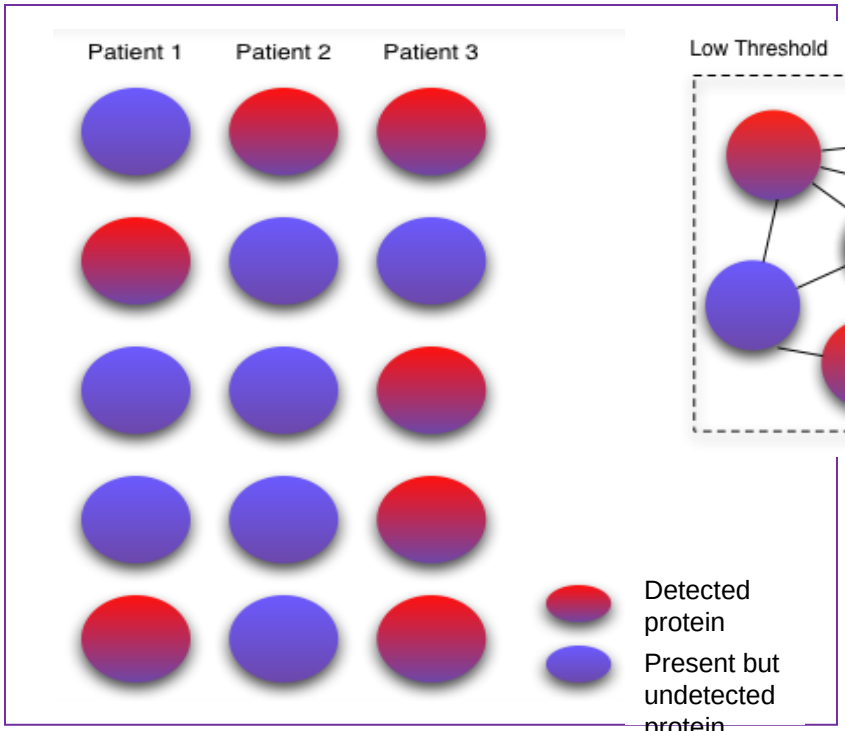
proteomic profile analysis

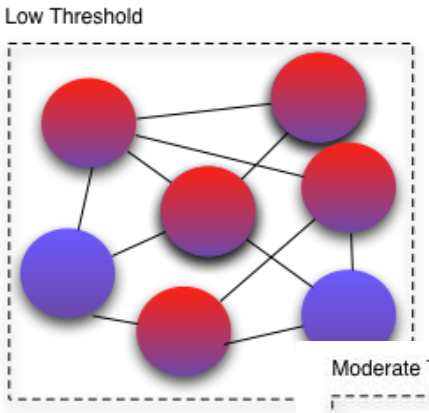
Issues in Proteomic Profiling

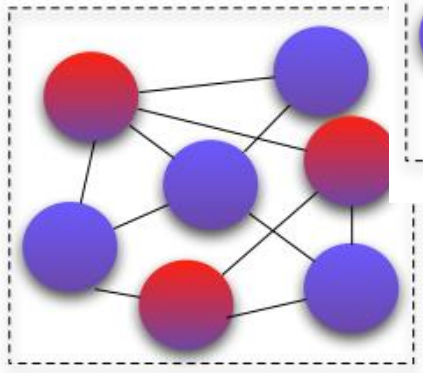
- **Coverage**
- **Consistency**

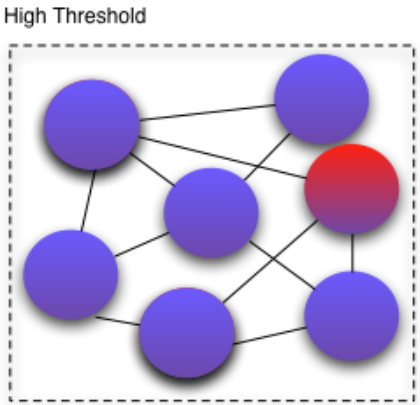
⇒ **Thresholding**

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

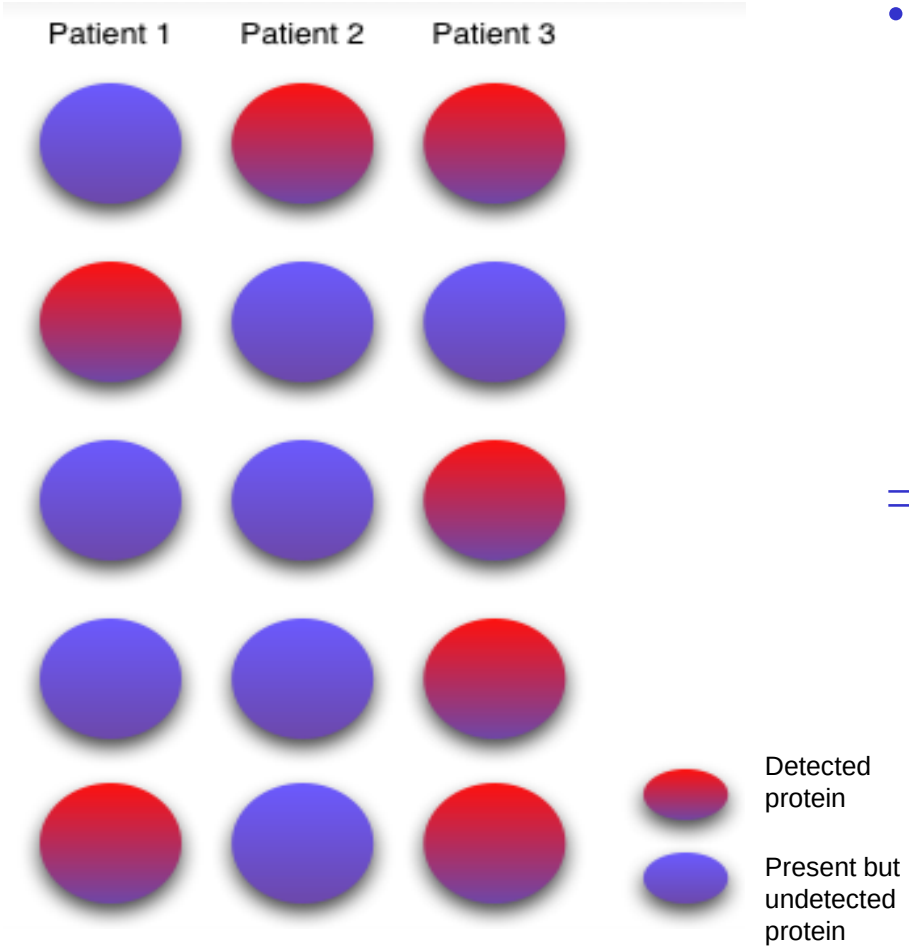








Intuitive Example

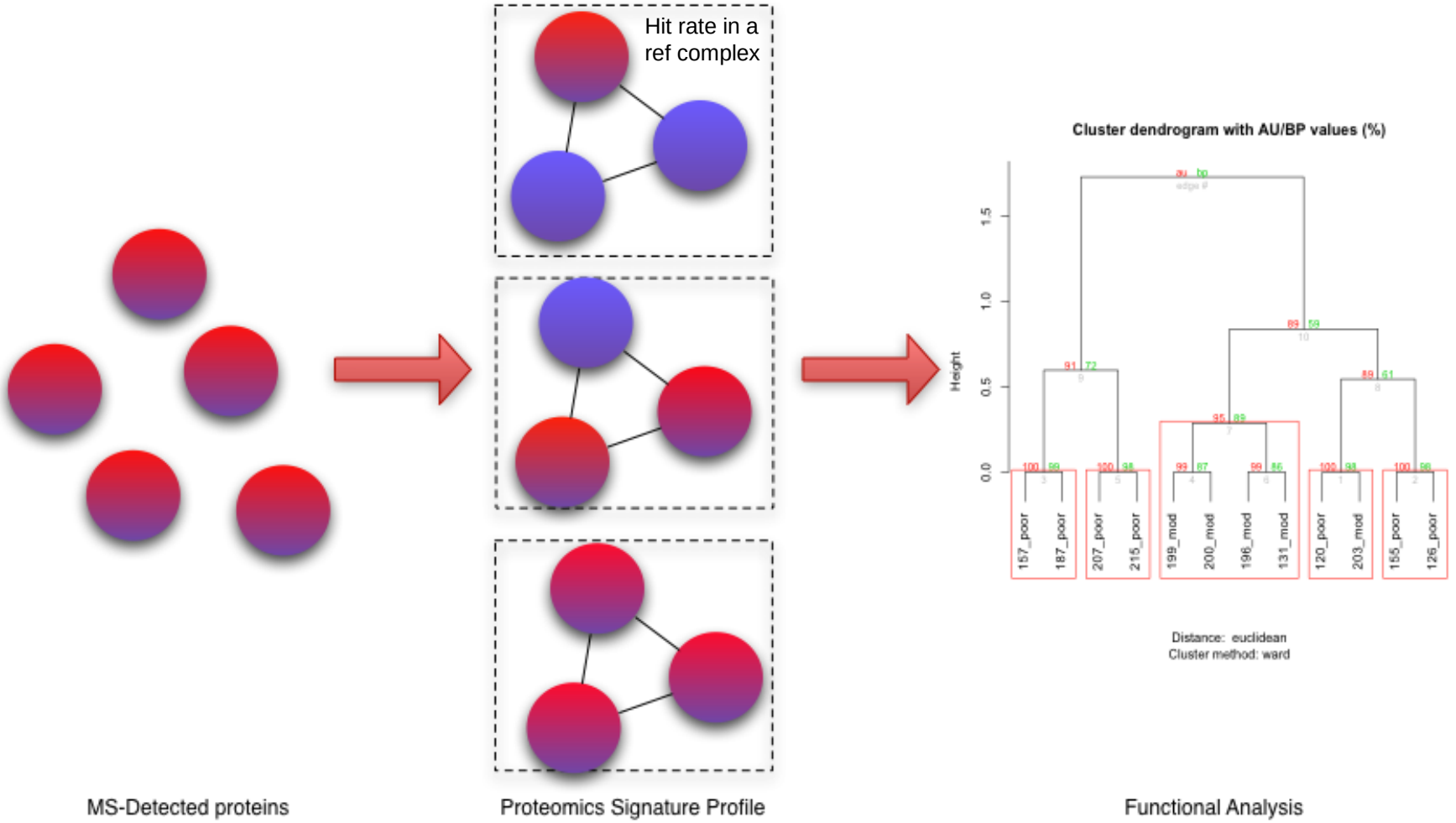


- **Suppose the failure to form a protein complex causes a disease**
  - If any component protein is missing, the complex can't form

⇒ **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**. *J.Proteome Research*, 11(3):1571-1581, 2012.

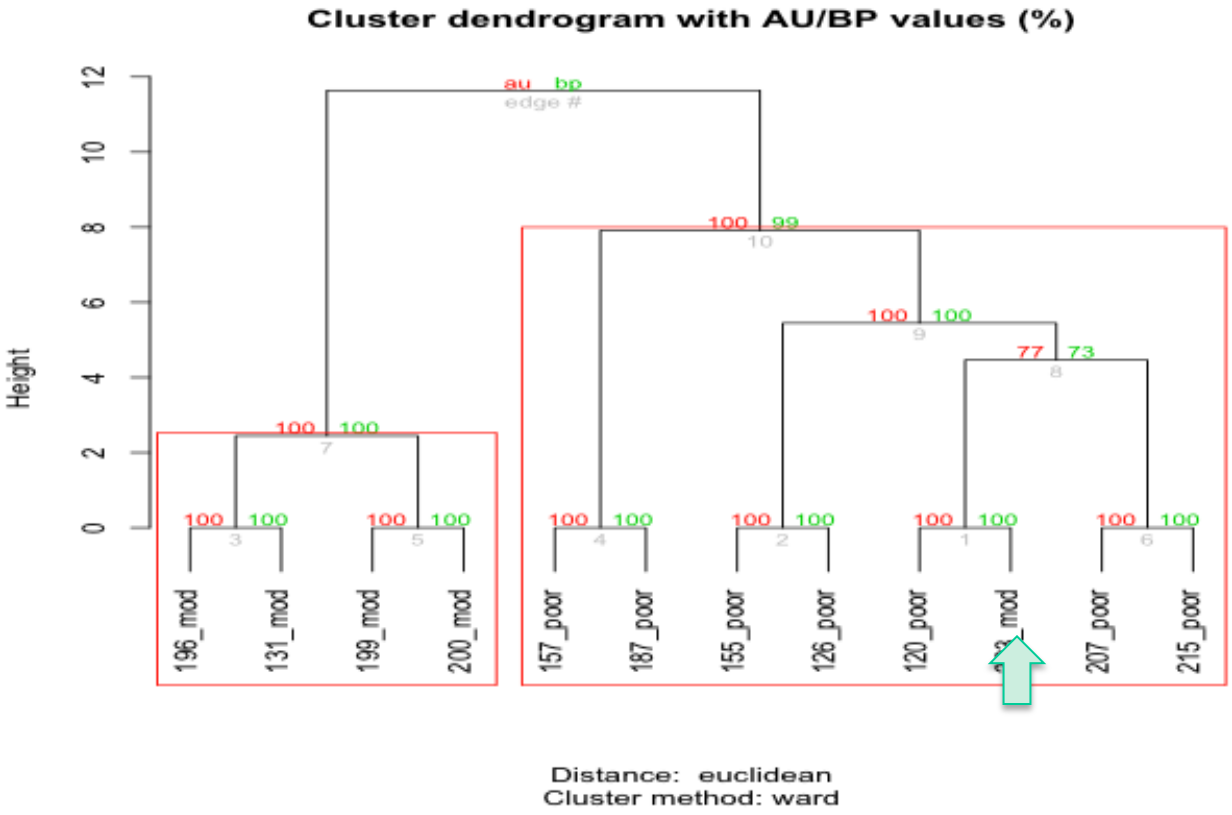
"Threshold-free" Principle of PSP



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

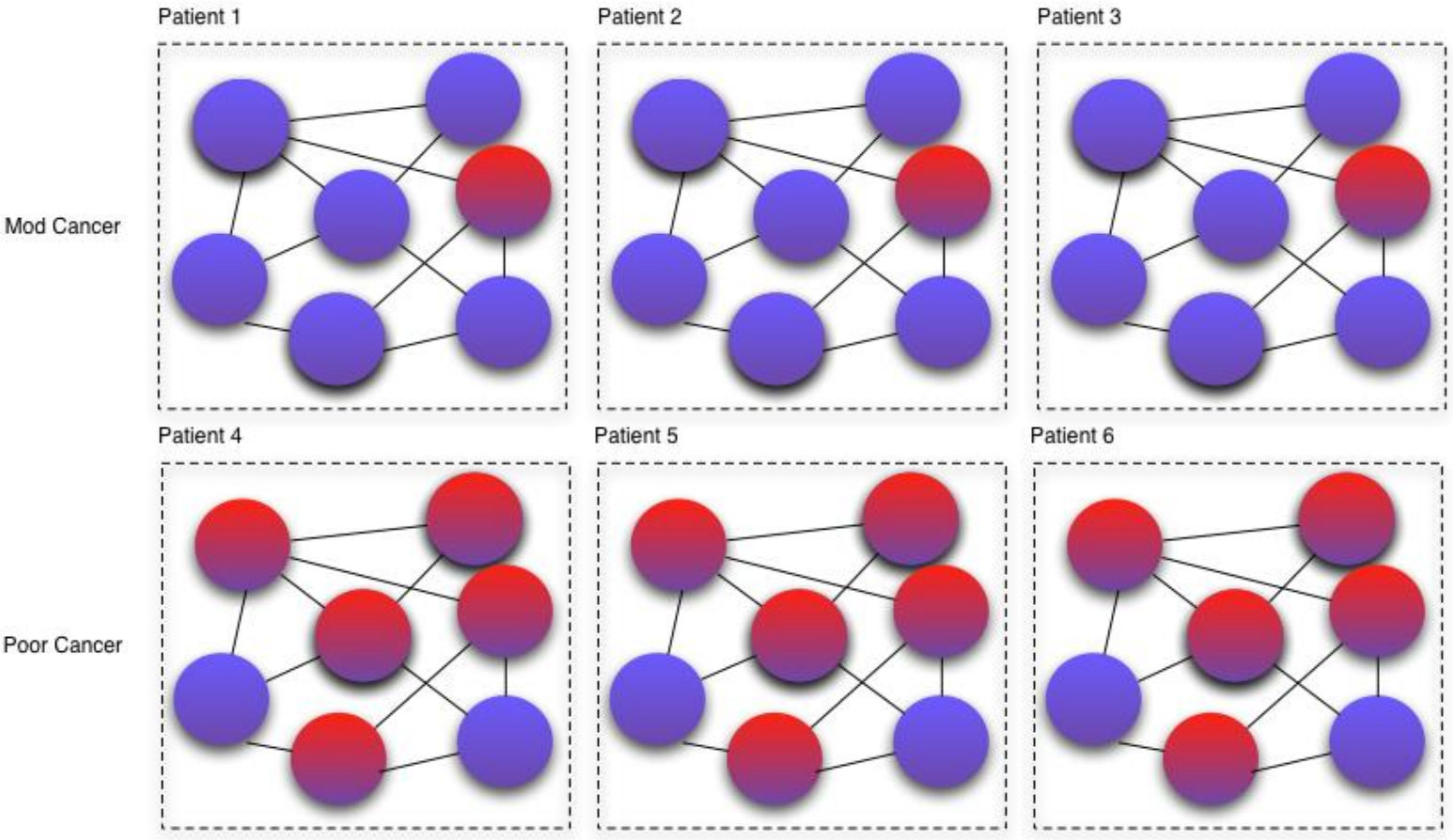
.

Consistency: Samples segregate by their classes with high confidence



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

Feature Selection



NUS National University of Singapore

concluding remarks

Conclusions

- **Consistent successful gene expression & proteomic profile analysis needs deep integration of background knowledge**

- **Most gene expression & proteomic profile analysis methods fail to give reproducible results when sample size is small (and some even fail when sample size is quite large)**

- **Logical analysis to identify key issues and simple logical solution to the issues can give fantastic results**

Acknowledgements

- **My students**
  - Kenny Chua, Kevin Lim, Dong Difeng
  - Donny Soh, Wilson Goh
  - Li Zhenhua
- **& collaborators**
  - Judy Sng, Maxey Chung, Choi Kwok Pui
- **Funding agencies**
  - MOE, A*STAR, NRF
  - Wellcome Trust

- Kevin Lim, Limsoon Wong. **Finding consistent disease subnetworks using PFSNet**. *Bioinformatics*, 30(2):189--196, 2014
- Kevin Lim, et al. **ESSNet: Finding consistent disease subnetworks in data with extremely small sample sizes**. In preparation
- Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *J Proteome Research*. 11(3):1571-1581, 2012
- Goh et al. **Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic acid-treated mice.** *J Proteome Research*, 12(5):2116-2127, 2013