Using Biological Networks for Protein Function Prediction, Biomarker Identification, and Other Problems in Computational Biology

Limsoon Wong

NUS
National University of Singapore

Outline of the Master Class

- **Brief overview of biological networks**

- **Using biological networks**
  - Gene expression profile analysis
  - Proteomic profile analysis
  - Protein function prediction
  - Other applications

- **Issues to be aware of in using biological networks**

Overview of Biological Networks

NUS
National University
of Singapore

Why Biological Networks?

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

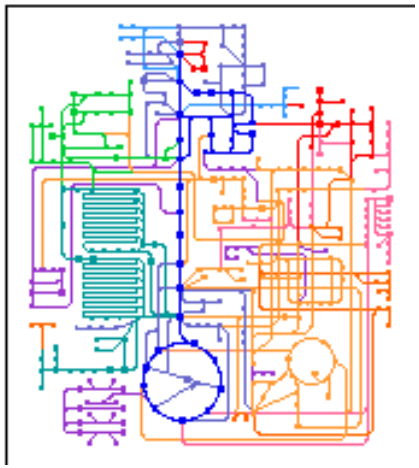
- **Proteins,** not genes, are responsible for many cellular activities

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

GENOME

PROTEOME

"INTERACTOME"



Slide credit: See-Kiong Ng

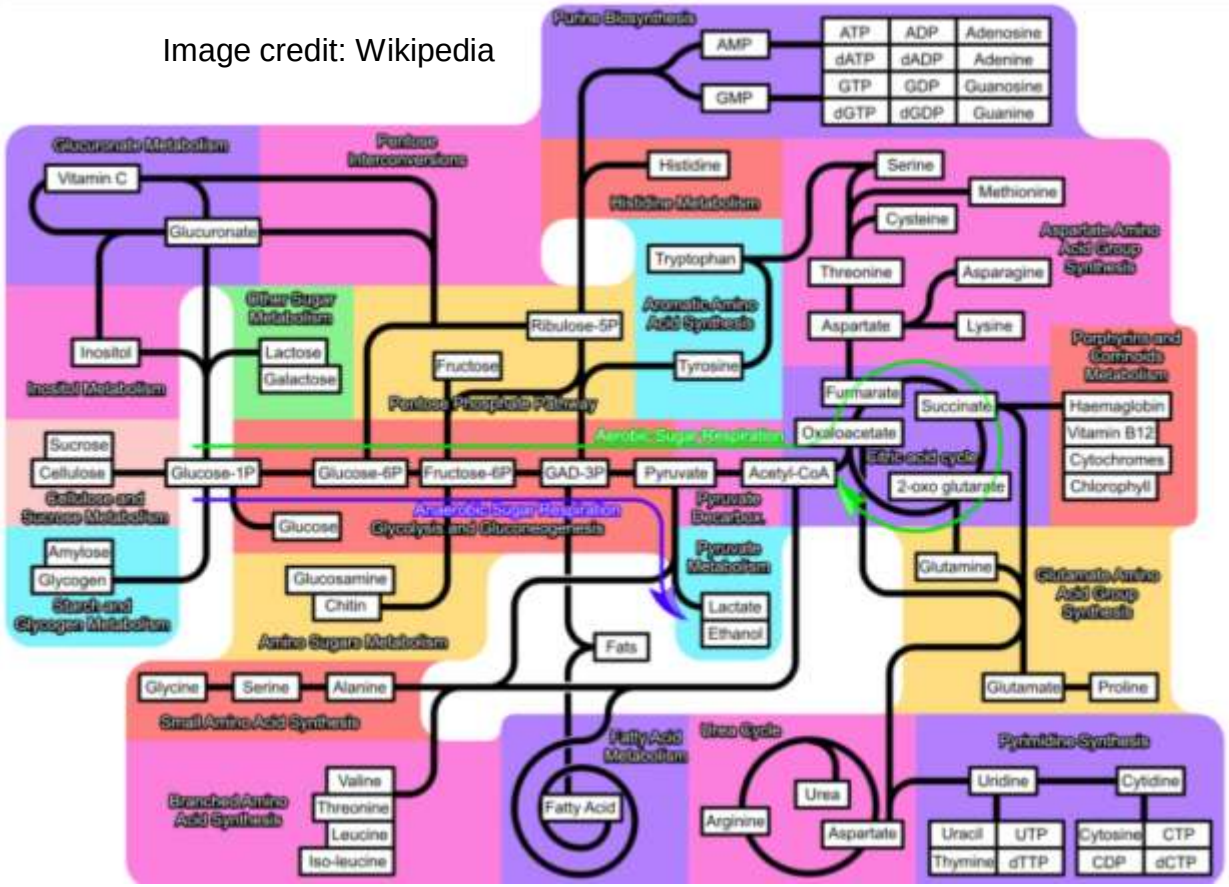
Types of Biological Networks

- **Natural biological pathways**
  - Metabolic pathway
  - Gene regulation network
  - Cell signaling network

- **Protein-protein interaction networks**

Metabolic Pathway

Image credit: Wikipedia



- **A series of biochem reactions in a cell**
  - Catalyzed by enzymes
  - Step-by-step modification of an initial molecule to form another product that can
    - **be used /store in the cell**
    - **initiate another metabolic pathway**

Gene Regulation Network

- **Gene regulation is the process that turns info from genes into gene products**
- **Gives a cell control over its structure & function**
  - Cell differentiation
  - Morphogenesis
  - Adaptability, …

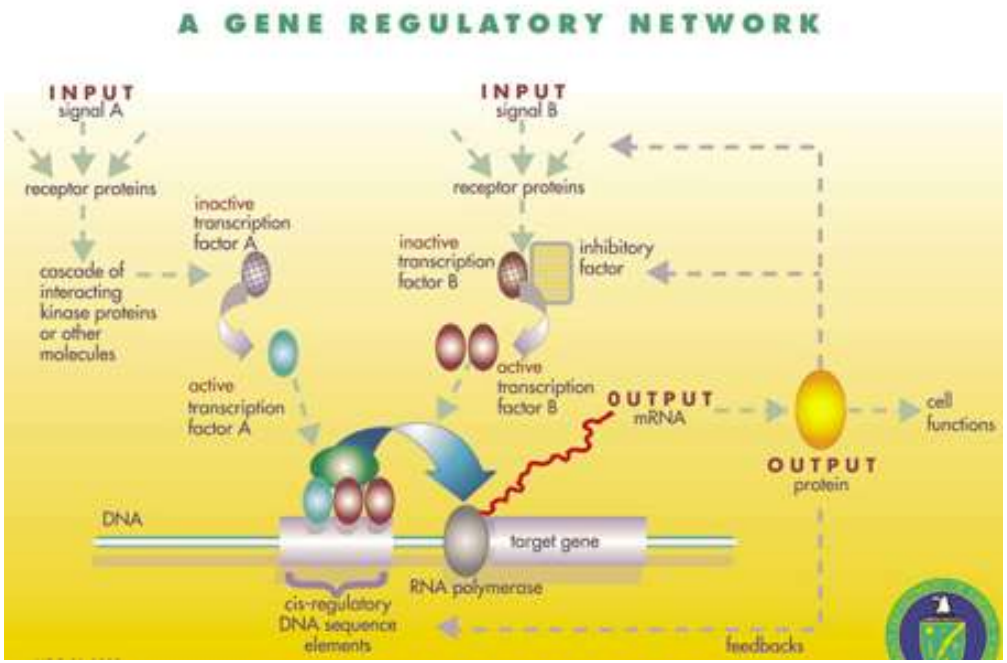


Image credit: Genome to Life

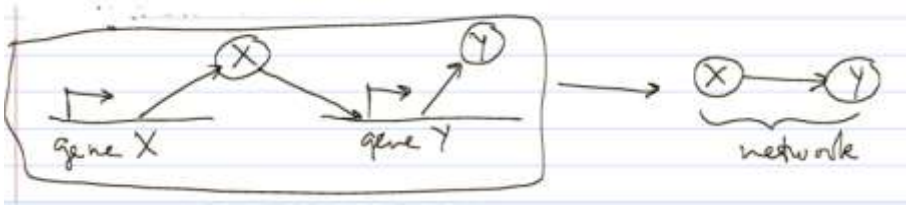


Image credit: Natasa Przulj

Cell Signaling Network

- **It is the entire set of changes induced by receptor activation**
  - Governs basic cellular activities and coordinates cell actions
- **Cells communicate with each other**
  - Direct contact (juxtacrine signaling)
  - Short distances (paracrine signaling)
  - Large distances (endocrine signaling)
- **Errors result in cancer, diabetes, ...**

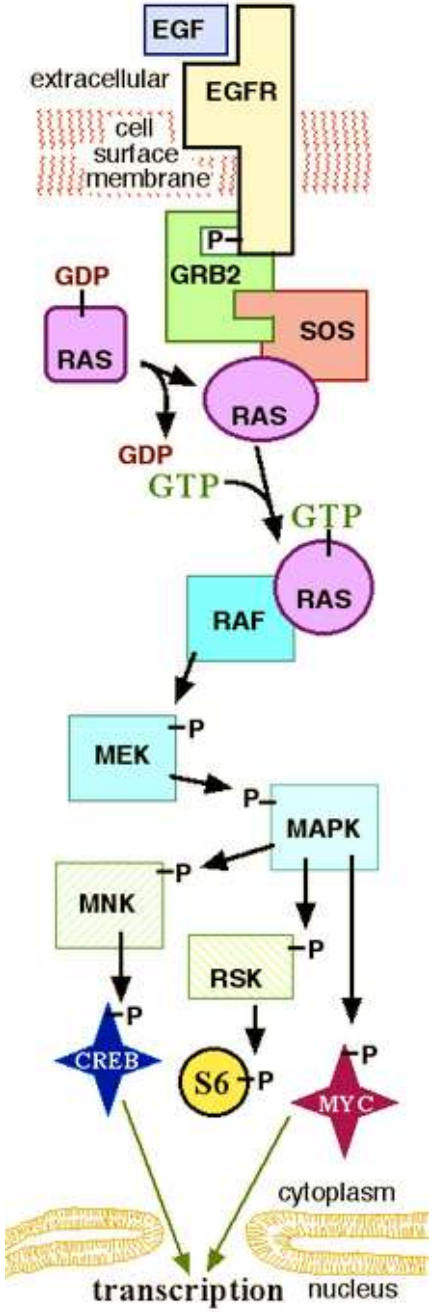
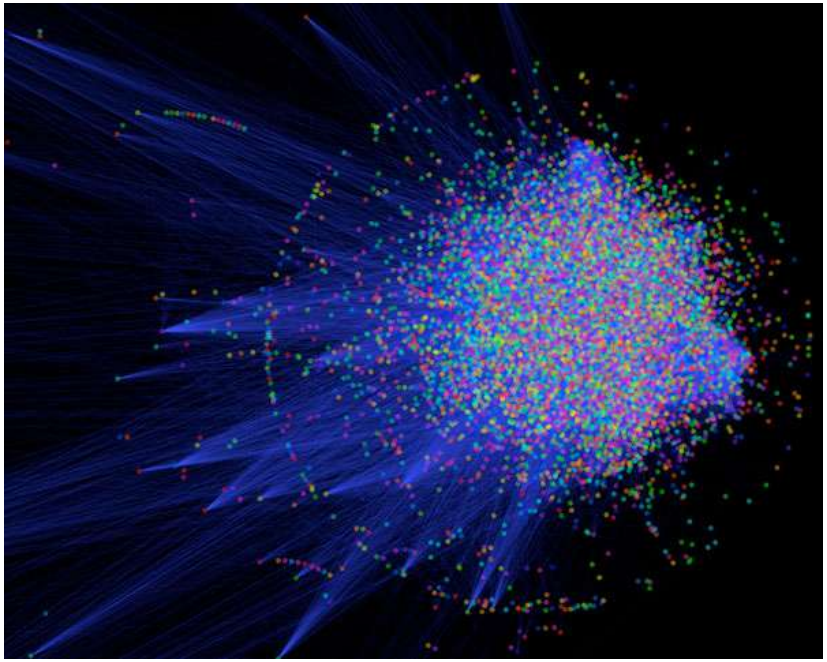


Image credit: Wikipedia

Protein Interaction Network (PPIN)

- **PPI usual refers to physical binding between proteins**
  - Stable interaction
    - **Protein complex**
    - **~70% of PPIs**
  - Transient interaction, modifying a protein for further actions
    - **Phosphorylation**
    - **Transportation**
    - **~30% of PPIs**



Visualization of the human interactome.
Image credit: Wikepedia

- **PPIN is usually a set of PPIs; it is not put into biological context**

Using Biological Networks, Part 1: *Delivering Reproducible Gene Expression Analysis*

Limsoon Wong

NUS
National University of Singapore

Part 1: Delivering reproducible gene expression analysis

- **Basic gene expression analysis**
- **Some issues in gene expression analysis**
- **Batch effect & normalization**
- **Improving reproducibility**

Gene Expression Measurement by Affymetrix GeneChip Array

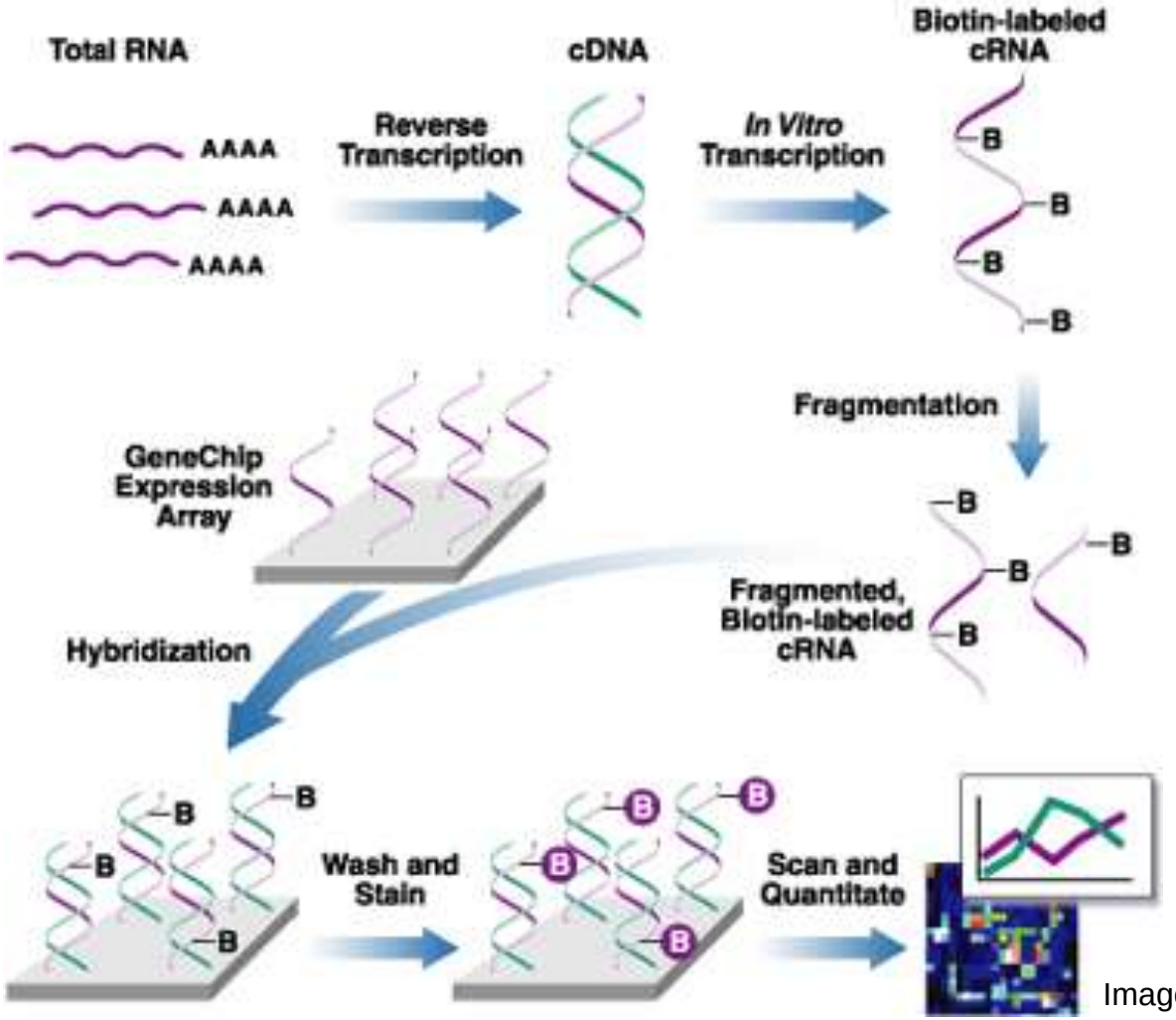
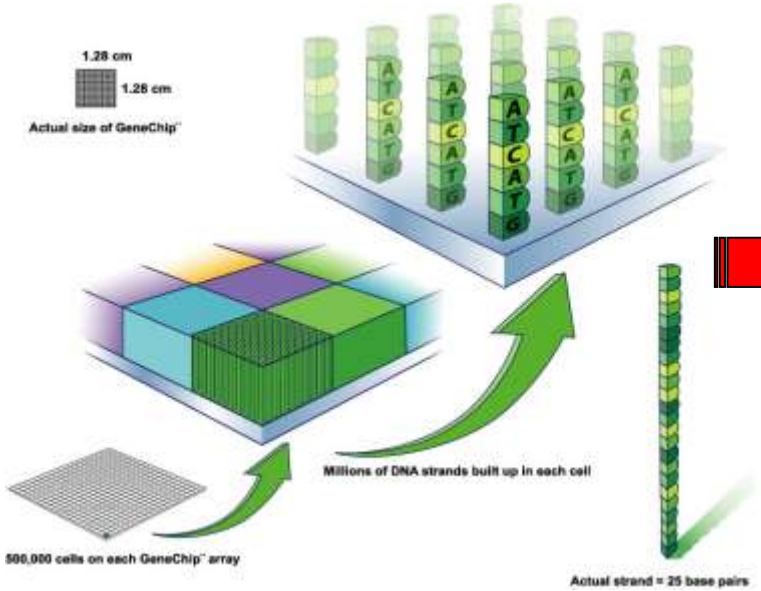


Image credit: Affymetrix

Diagnosis Using Microarray



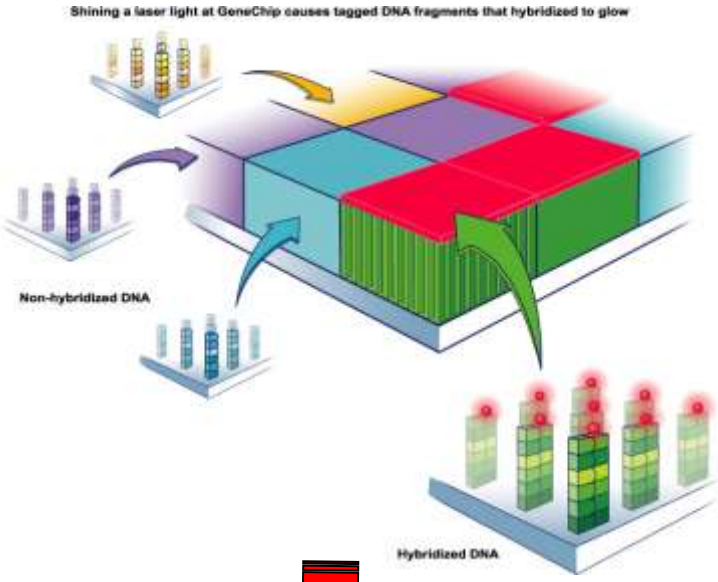
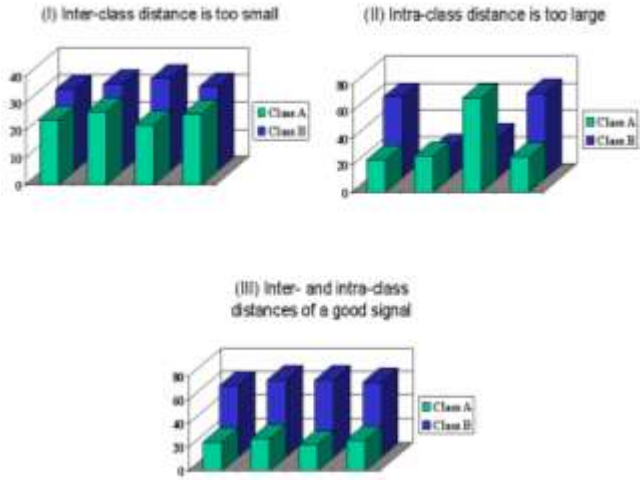
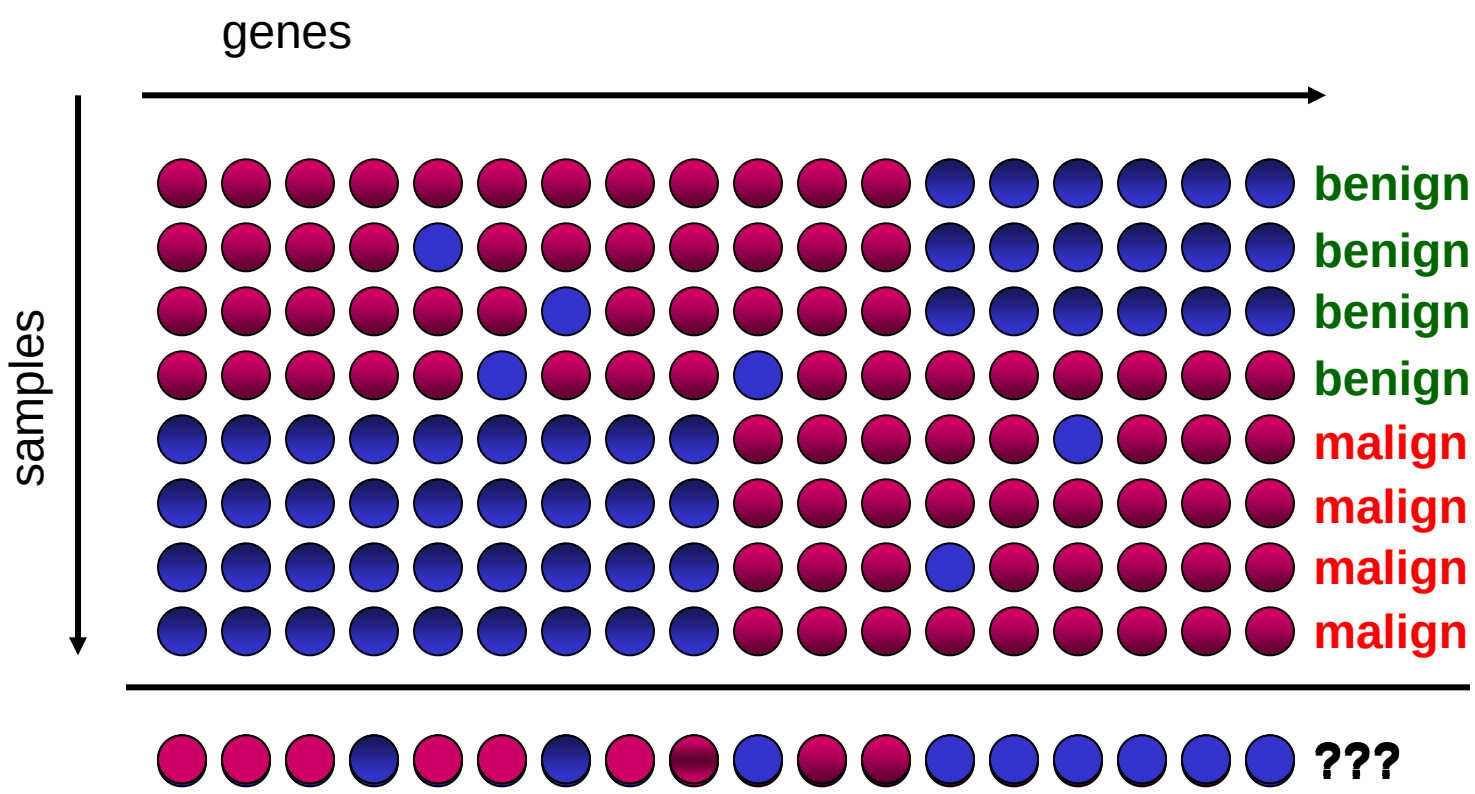


Image credit: Affymetrix

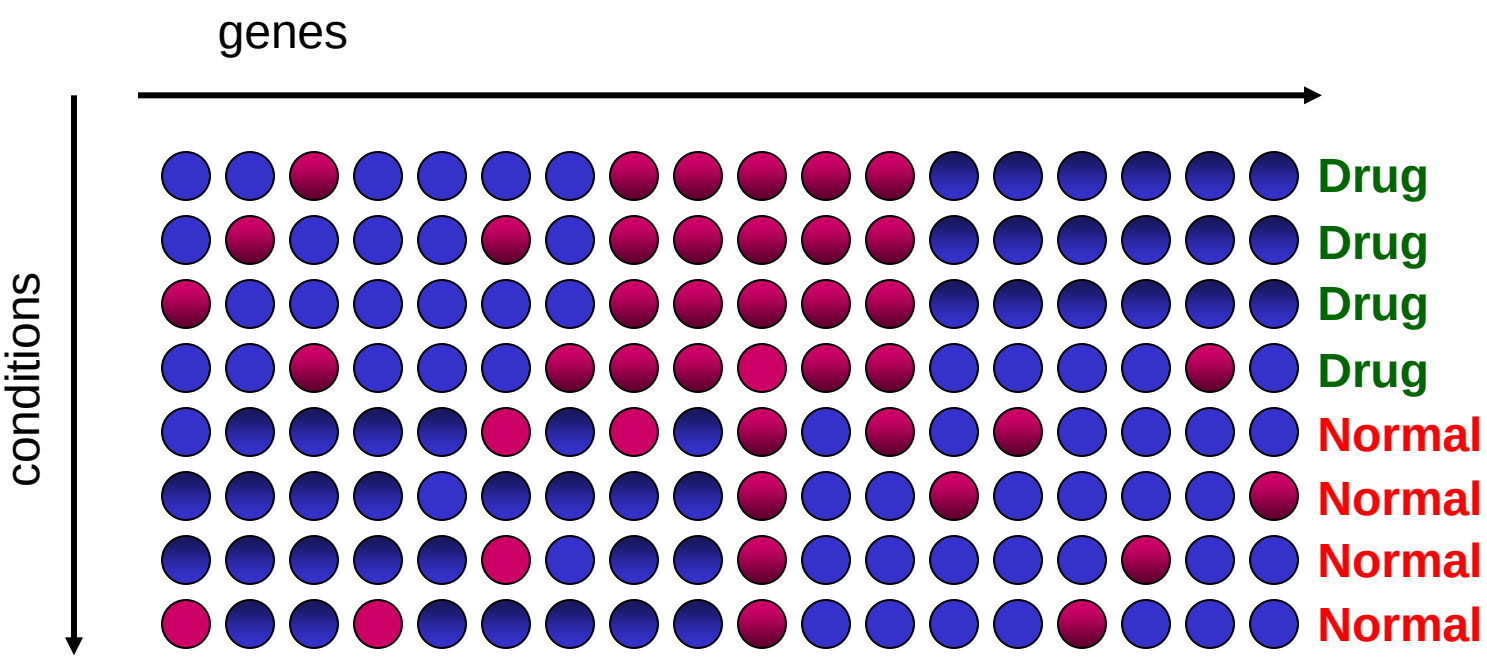


	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAv	Avg Diff	Abs Call	
AFFX-Murl	5	2	19	297.5	A	M16762 Mouse int
AFFX-Murl	3	2	19	554.2	A	M37897 Mouse int
AFFX-Murl	4	2	19	308.6	A	M25892 Mus mus
AFFX-Murf	1	3	19	141	A	M83649 Mus mus
AFFX-BioE	13	1	19	9340.6	P	J04423 E coli bioE
AFFX-BioE	15	0	19	12862.4	P	J04423 E coli bioE
AFFX-BioE	12	0	19	8716.5	P	J04423 E coli bioE
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacteriop
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacteriop
AFFX-BioE	7	5	18	-483	A	J04423 E coli bioE
AFFX-BioE	5	4	18	313.7	A	J04423 E coli bioE
AFFX-BioE	7	6	20	-1016.2	A	J04423 E coli bioE

Application: Disease Subtype Diagnosis



Application: Drug Action Detection



Which group of genes are the drug affecting on?

Typical Analysis Workflow

- **Gene expression data collection**
- **DE gene selection by, e.g., t-statistic**
- **Classifier training based on selected DE genes**
- **Apply the classifier for diagnosis of future cases**

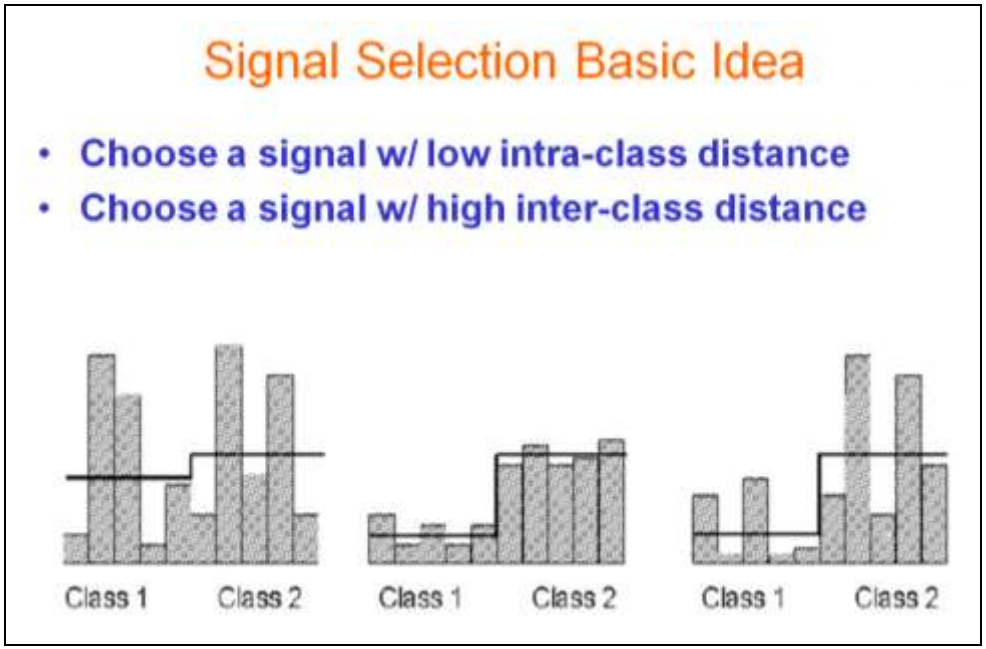


Image credit: Golub et al., *Science*, 286:531–537, 1999

Hierarchical Clustering

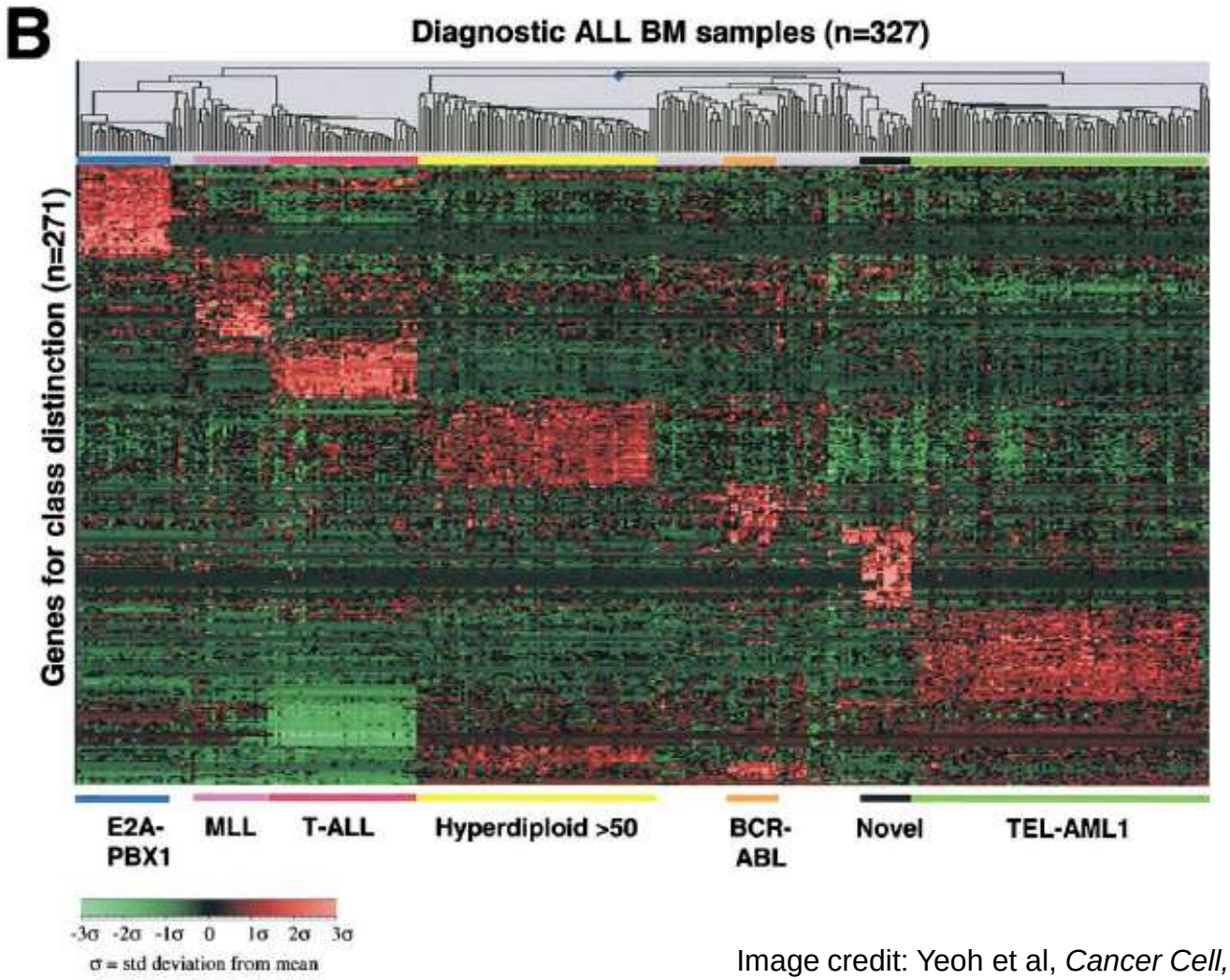
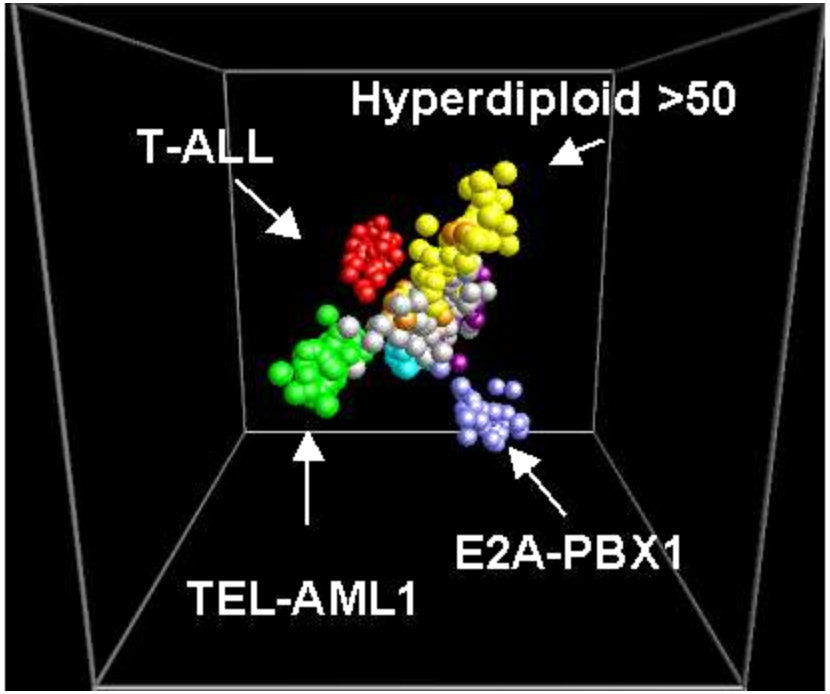


Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

PCA Plots



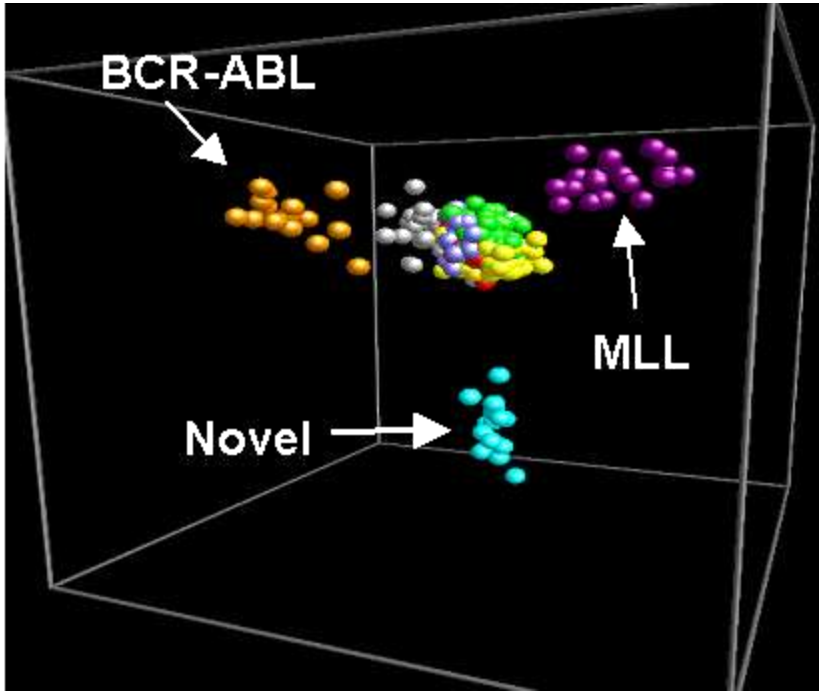


Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

Part 1: Delivering reproducible gene expression analysis

- Basic gene expression analysis
- **Some issues in gene expression analysis**
- Batch effect & normalization
- Improving reproducibility

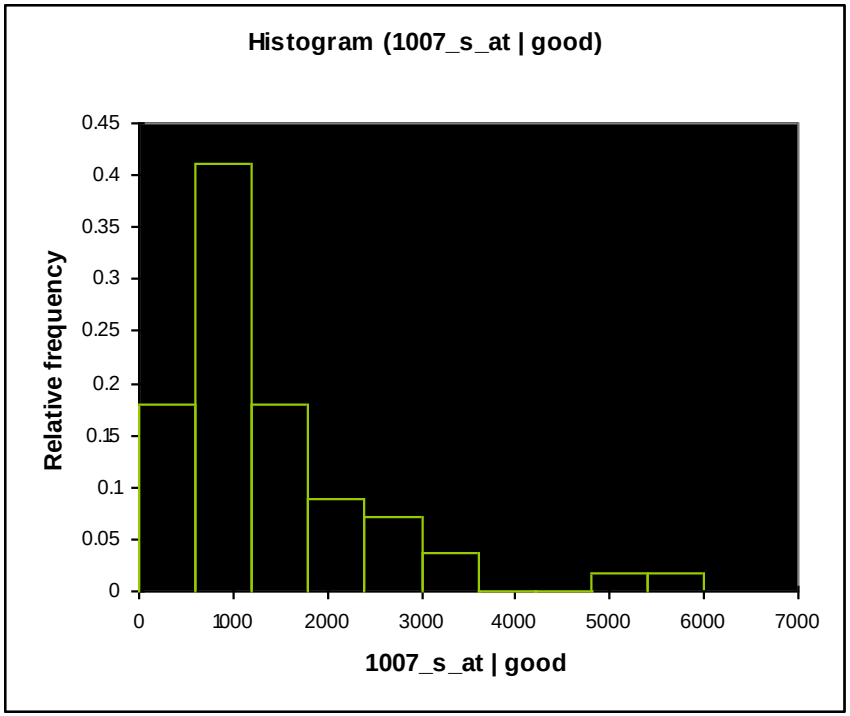
Some Headaches

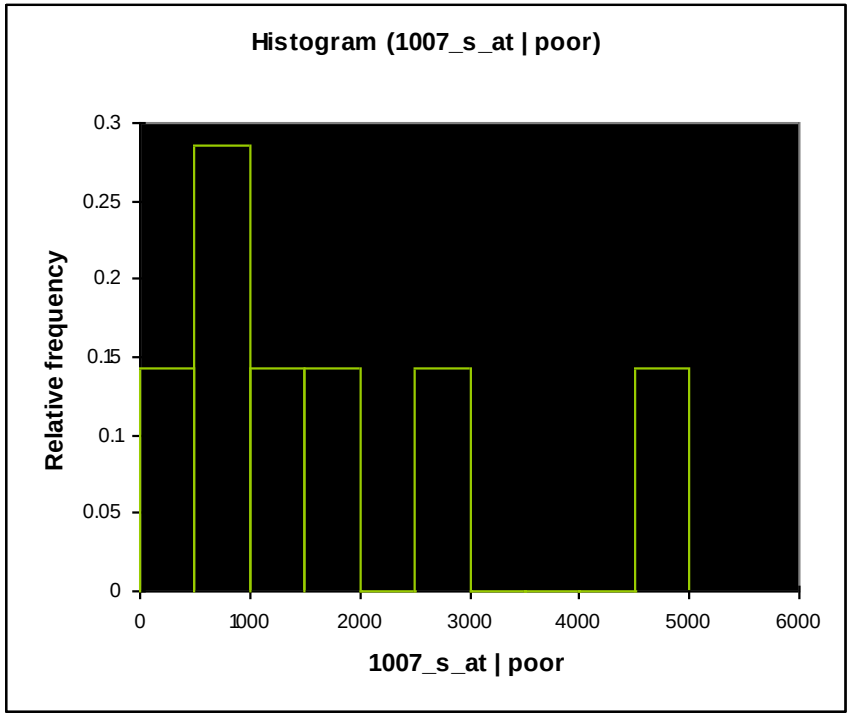
- **Natural fluctuations of gene expression in a person**

- **Noise in experimental protocols**
  - Numbers mean diff things in diff batches
  - Numbers mean diff things in data obtained from diff platforms

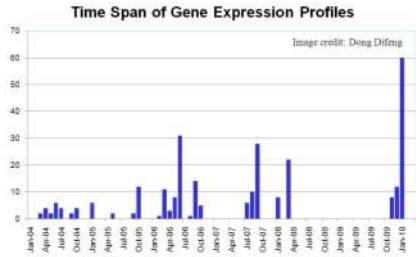
⇒ **Selected genes may not be meaningful**
  - Diff genes get selected in diff expts

Natural Fluctuations

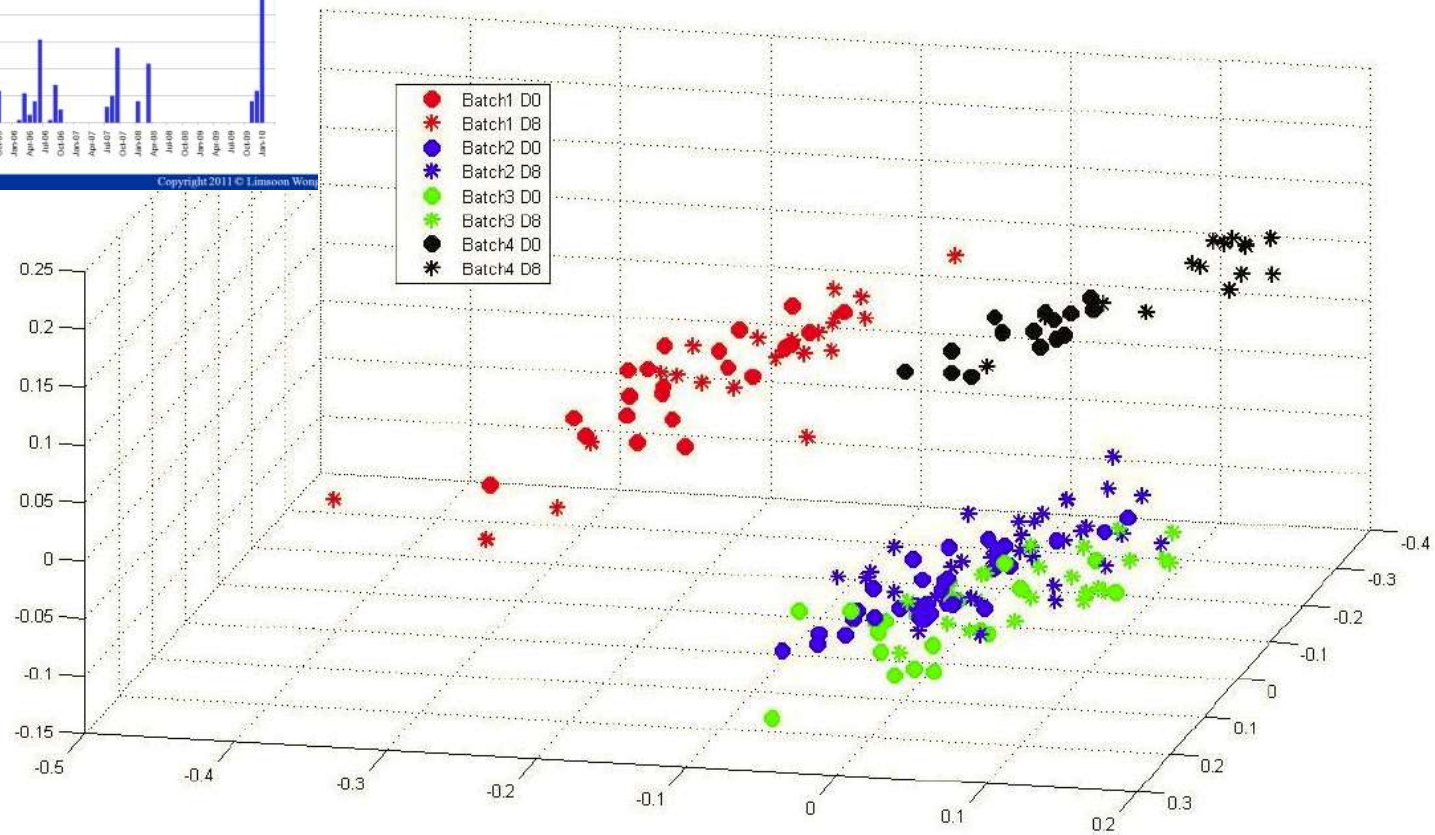




Sometimes, a gene expression study may involve batches of data collected over a long period of time…



Batch Effects



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

Image credit: Difeng Dong's PhD dissertation, 2011

Percentage of Overlapping Genes

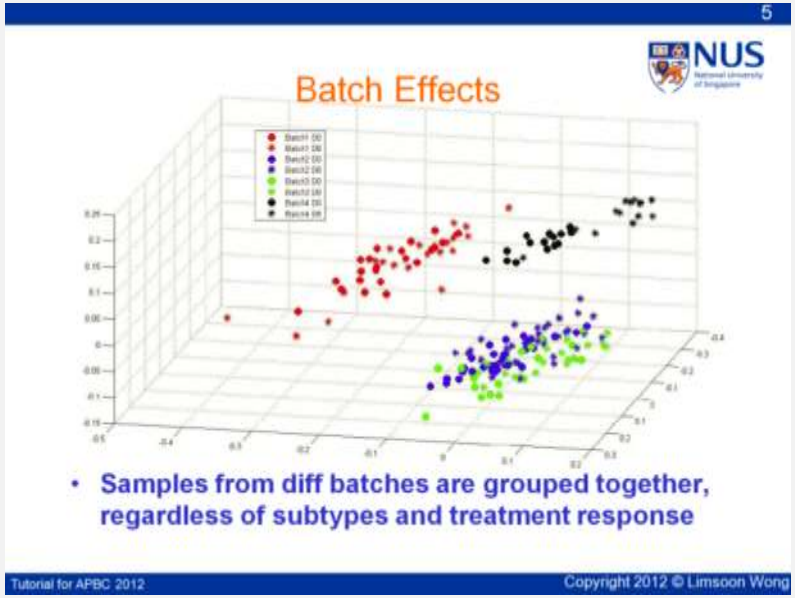
- **Low % of overlapping genes from diff expt in general**
  - Prostate cancer
    - **Lapointe et al, 2004**
    - **Singh et al, 2002**
  - Lung cancer
    - **Garber et al, 2001**
    - **Bhattacharjee et al, 2001**
  - DMD
    - **Haslett et al, 2002**
    - **Pescatori et al, 2007**

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

Part 1: Delivering reproducible gene expression analysis

- Basic gene expression analysis
- Some issues in gene expression analysis
- **Batch effect & normalization**
- Improving reproducibility



Approaches to Normalization

- **Aim of normalization:**
  Reduce variance w/o increasing bias

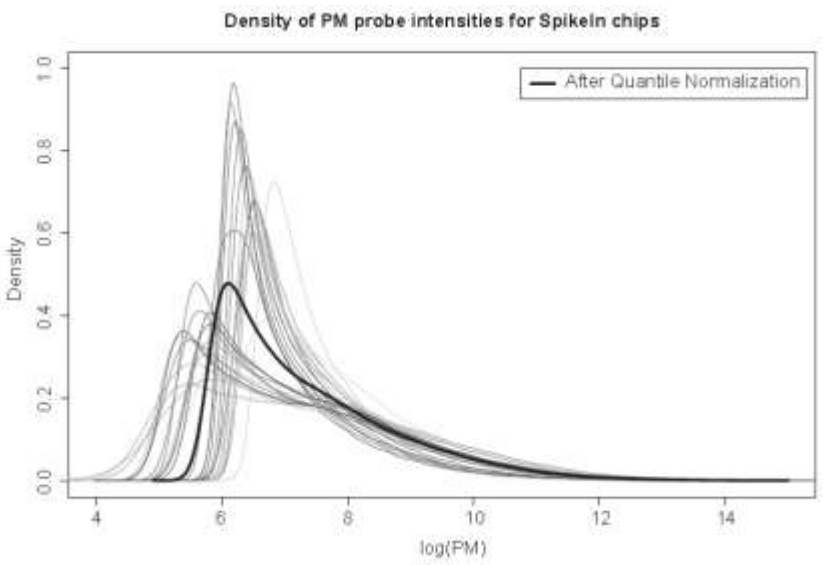
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

- **Transform data so that distribution of probe intensities is same on all arrays**
  - E.g., $(x - \mu) / \sigma$

- **Quantile normalization**

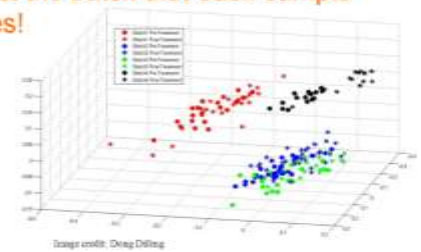
Quantile Normalization

- **Given *n arrays of length p, form X of size p × n where each array is a column***
- **Sort each column of *X to give* X_{sort}**
- **Take means across rows of X_{sort} *and assign this* mean to each elem in the row to get X'_{sort}**
- **Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as *X***



- Implemented in some microarray s/w, e.g., EXPANDER

In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!



⇒ Need normalization to correct for batch effect

After quantile normalization

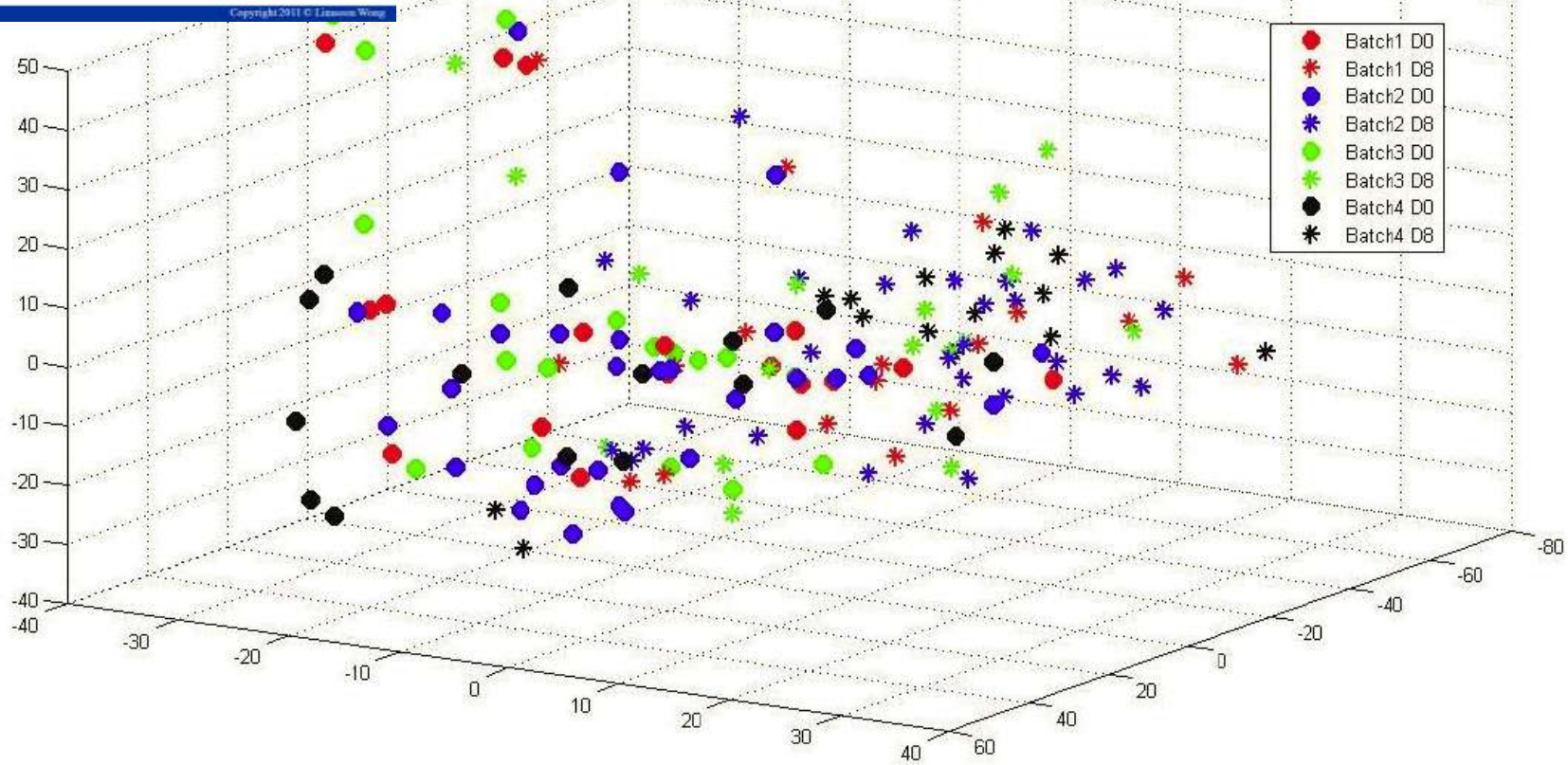
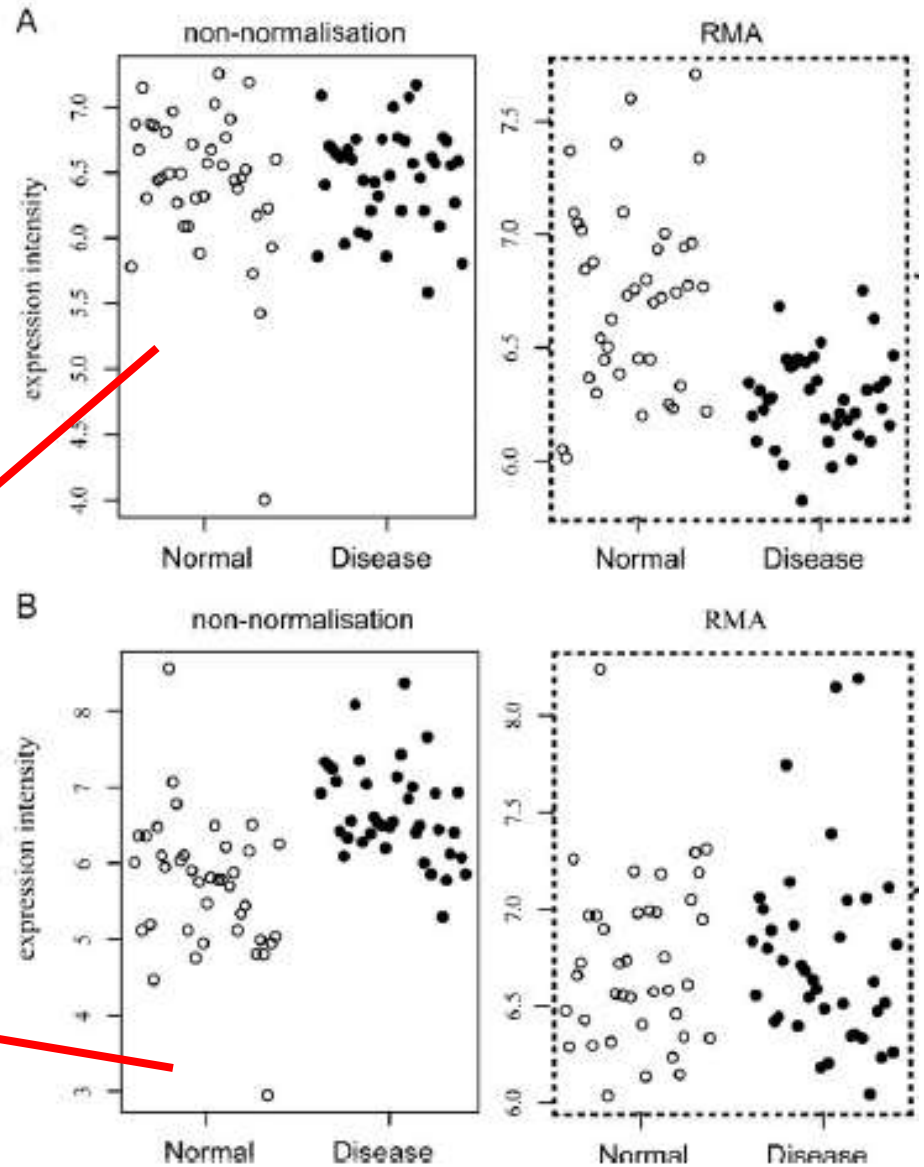


Image credit: Difeng Dong's PhD dissertation, 2011

Caution: “Over normalize” signals in cancer samples

A gene normalized by quantile normalization (RMA) was detected as down-regulated DE gene, but the original probe intensities in cancer samples were higher than those in normal samples

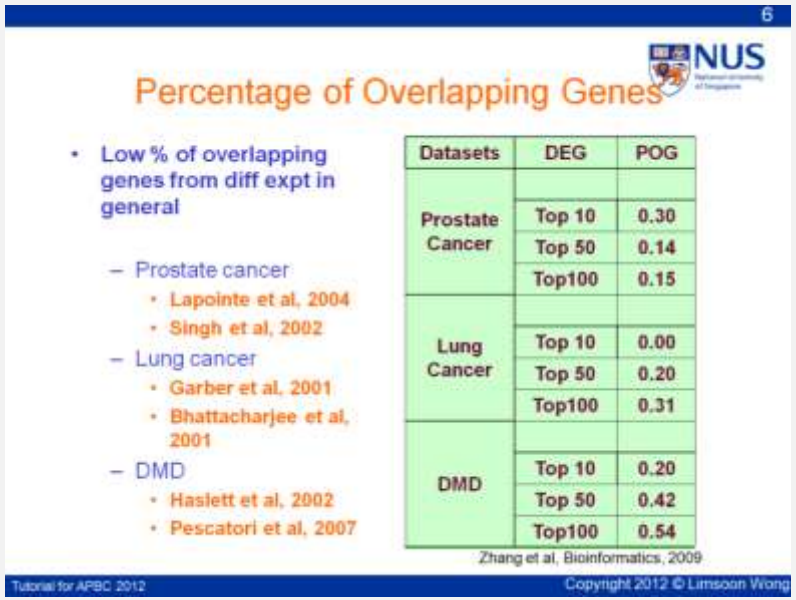
A gene was detected as an up-regulated DE gene in the non-normalized data, but was not identified as a DE gene in the quantile nornmalized data



Wang et al. *Molecular Biosystems*, in press

Part 1: Delivering reproducible gene expression analysis

- Basic gene expression analysis
- Some issues in gene expression analysis
- Batch effect & normalization
- **Improving reproducibility**



6

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

Tutorial for APBC 2012 Copyright 2012 © Limsoon Wong

Law of Large Numbers

- Suppose you are in a room with 365 other people
- Q: What is prob that a specific person in the room has the same birthday as you?
- A: 1/365 = 0.3%
- Q: What is prob that there is a person in the room having same birthday as you?
- A: $1 - (364/365)^{365} = 63\%$
- Q: What is prob that there are two persons in the room having same birthday?
- A: 100%

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**

$\Rightarrow$ **E(# of correlated genes) = 1,562**

$\Rightarrow$ **Many false positives**

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

Group of Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **What is the chance of a group of 5 genes being perfectly correlated to these samples?**

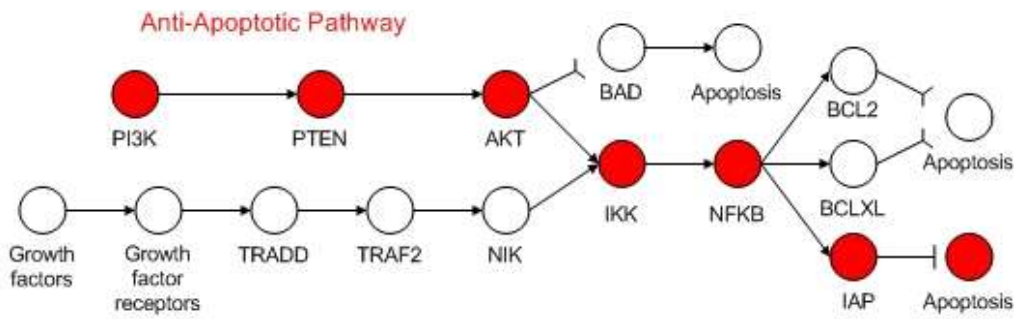
- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- **# of groups = ${}^{100000}C_5$**

$\Rightarrow$ **E(# of groups of genes correlated) = ${}^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$**

$\Rightarrow$ **Even more false positives?**

- **Perhaps no need to consider every group**

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**

Taming false positives by considering pathways instead of all possible groups

Group of Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **What is the chance of a group of 5 genes being perfectly correlated to these samples?**

- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- ~~**# of groups = $^{100000}C_5$**~~
- ~~**E(# of groups of genes correlated) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$**~~

⇒ **Even more false positives?**

- **Perhaps no need to consider every group**

Microarray Workshop for Gene Expression Profiling, NUH, 23/9/2011 Copyright 2011 © Limsoon Wong

# of pathways = 1000

E(# of pathways correlated) = $1000 * (1/2^6)^5 = 9.3*10^{-7}$

Towards More Meaningful Genes

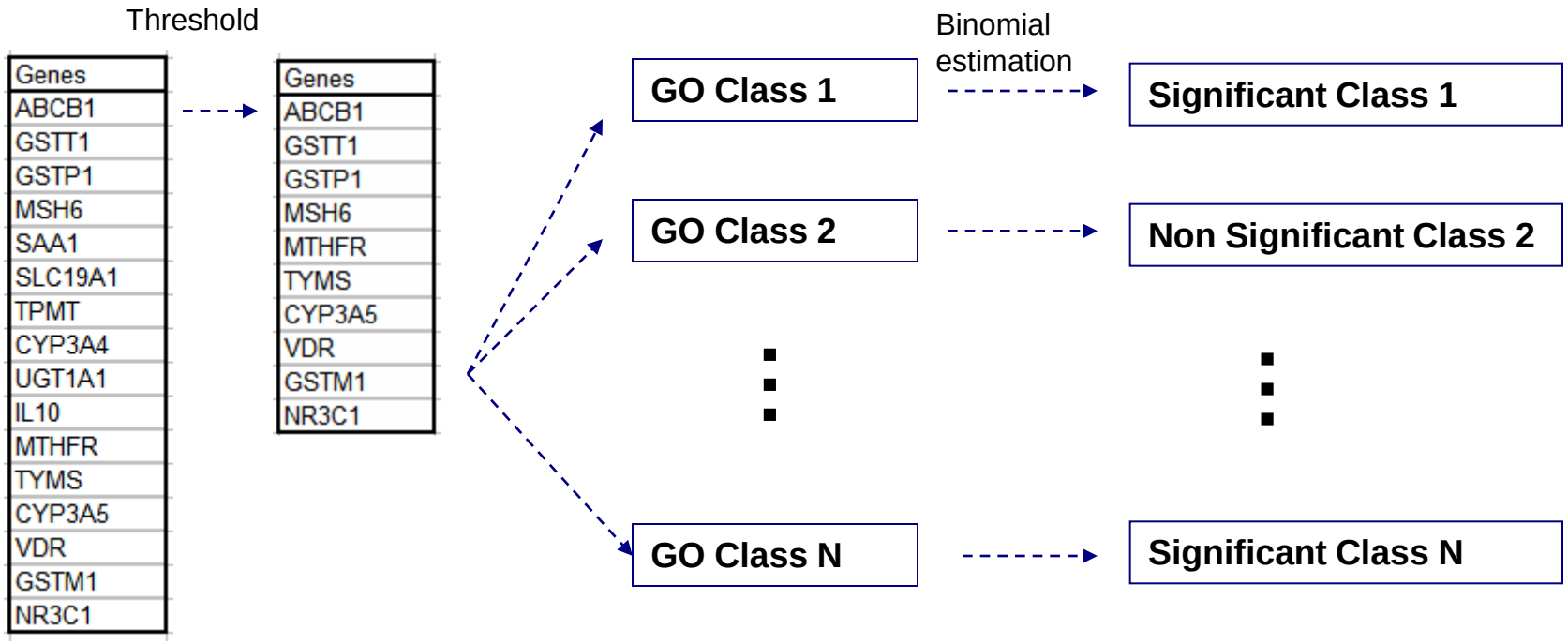
- **ORA**
  - Khatri et al
  - *Genomics*, 2002
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - *PNAS*, 2005
- **SNet**
  - Soh et al
  - *BMC Genomics*, 2011

Overlap Analysis

Direct-Group Analysis

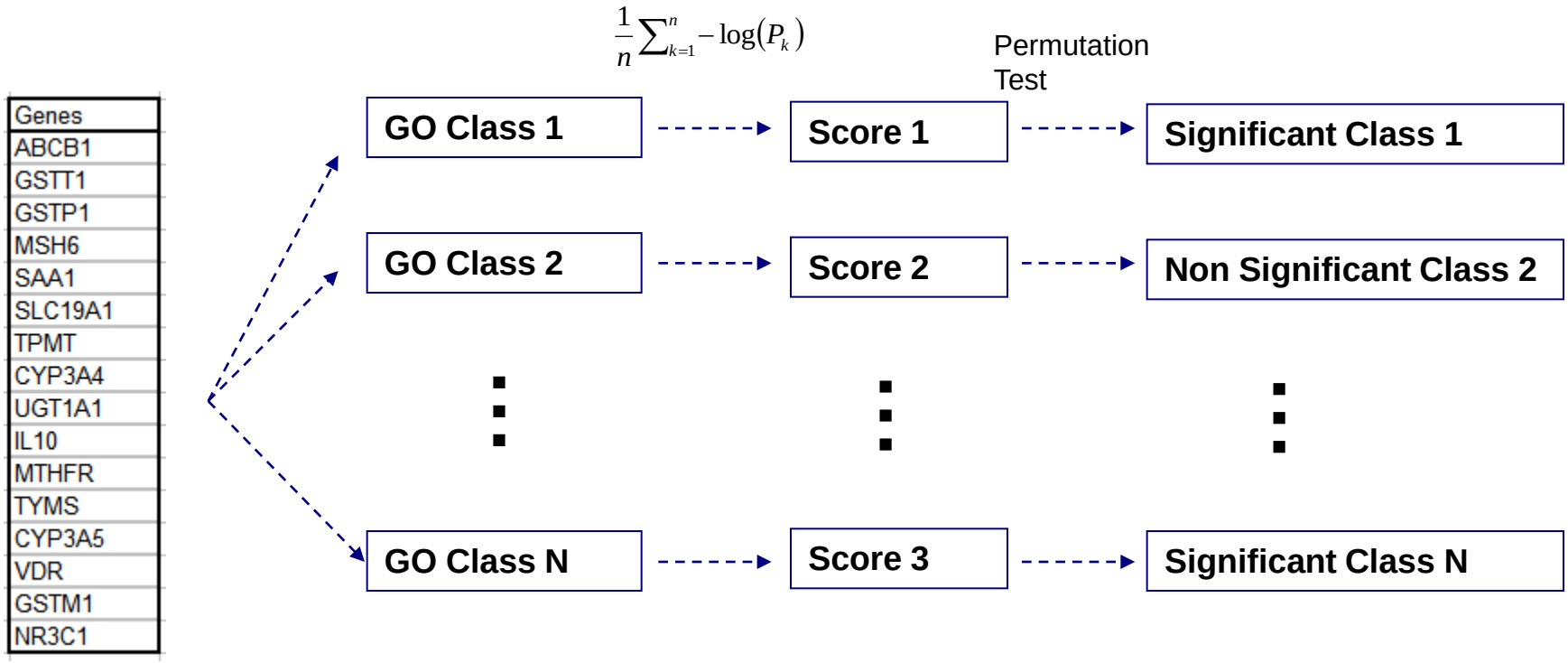
Network-Based Analysis

Overlap Analysis: ORA



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

Direct-Group Analysis: FCS

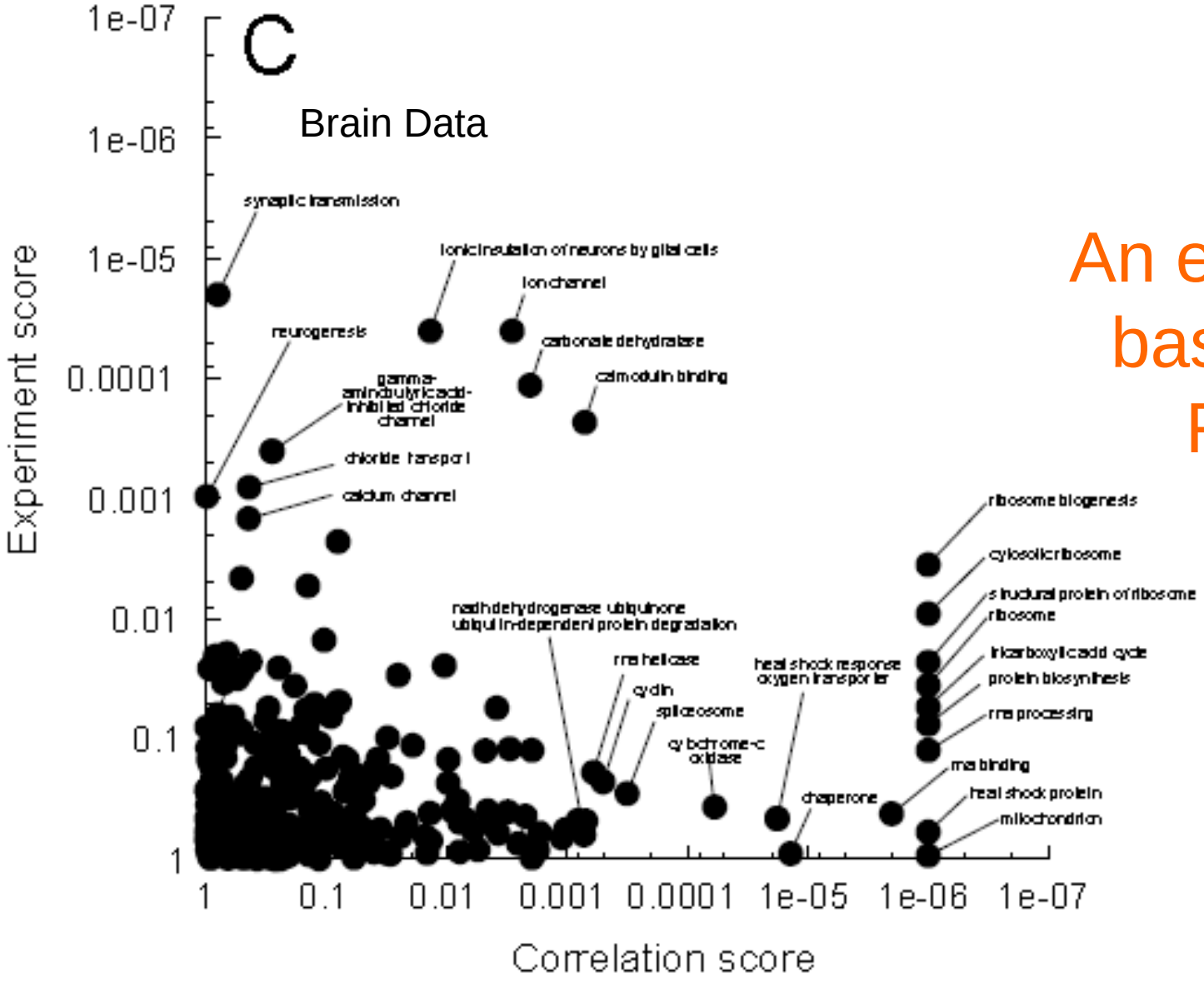


P Pavlidis et al. "Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex". *Neurochem Res*., 29(6):1213-1222, 2004.

FCS: Key variations

- **"Correlation score"**
  - Score of a class C = average pair-wise correlation of genes in the class C
- **"Experimental score"**
  - Score of a class C = average of log-transformed p-values of genes in the class C
- **Null distribution to estimate the p-value of the scores above is by repeated sampling of random sets of genes of the same size as C**

Pavlidis et al., *PSB 2002*



An example based on FCS

Pavlidis et al., *PSB 2002*

Goeman & Buhlmann. "Analyzing gene expression data in terms of gene sets: Methodological issues". *Bioinformatics*, 23(8):980-987, 2007

A problem w/ FCS as proposed by Pavlidis et al in PSB 2002

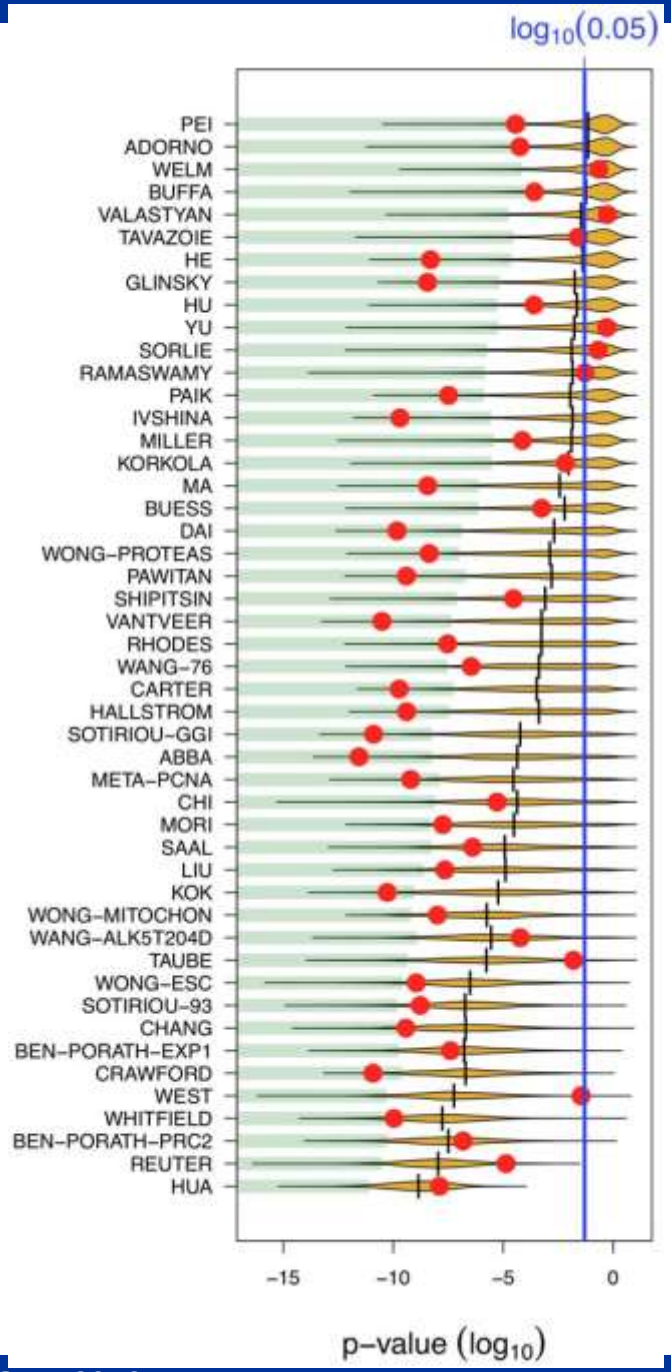
FCS: Key variations

- **"Correlation score"**
  - Score of a class C = average pair-wise correlation of genes in the class C
- **"Experimental score"**
  - Score of a class C = average of log-transformed p-values of genes in the class C
- **Null distribution to estimate the p-value of the scores above is by repeated sampling of random sets of genes of the same size as C**

Pavlidis et al., PSB 2002

- **Its null hypothesis:**
  - "genes in C are independently expressed & not diff from other genes
- **But …**
  - Genes in a pathway are not independent

  ⇒ Becomes over sensitive
- **Solution: generate null distribution by randomizing patient class labels**



FCS: Why do we estimate p-value using a null distribution based on repeated sampling of randomized gene sets / patient sets?

Venet et al. “Most random gene expression signatures are significantly associated with breast cancer outcome”. *PLoS Computational Biology*, 7(10):e1002240, 2011.

Direct-Group Analysis: GSEA

Rank Genes

Assign score to each class based on gene rank

Permutation test

$P_{hit}(S_1, i)$

Genes	p-values
ABCB1	0.0040
GSTT1	0.0051
GSTP1	0.0126
MSH6	0.0135
SAA1	0.0386
SLC19A1	0.0410
TPMT	0.0423
CYP3A4	0.0500
UGT1A1	0.0610
IL10	0.0626
MTHFR	0.0756
TYMS	0.0871
CYP3A5	0.0879
VDR	0.0906
GSTM1	0.0949
NR3C1	0.0991

GO Class 1 → **Significant Class 1**

$\max(P_{hit}(S_1, i) - P_{miss}(S_1, i))$

$P_{miss}(S_1, i)$

GO Class 2 → **Non Significant Class 2**

$P_{hit}(S_2, i)$ $P_{miss}(S_2, i)$

⋮ ⋮

GO Class N → **Significant Class N**

$P_{hit}(S_N, i)$ $P_{miss}(S_N, i)$

Subramanian et al. “Gene set enrichment analysis: A knowledge-based approach for interpreting genome wide expression profiles”. *PNAS*, 102(43):15545-15550, 2005

GSEA: Key Points

- **"Enrichment score"**
  - The degree that the genes in gene set C are enriched in the extremes of ranked list of all genes
  - Measured by Komogorov-Smirnov statistic

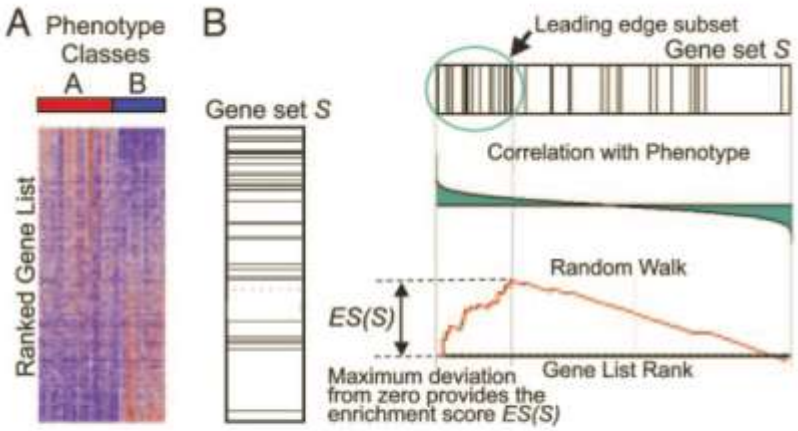


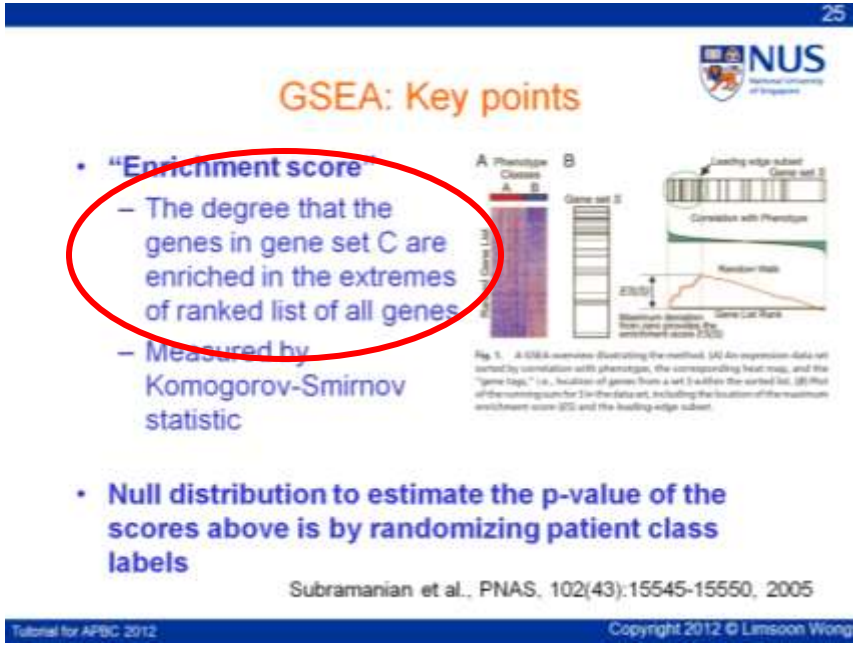
Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Subramanian et al., *PNAS*, 102(43):15545-15550, 2005

- **Null distribution to estimate the p-value of the scores above is by randomizing patient class labels**

Wong. "Using Biological Networks in Protein Function Prediction and Gene Expression Analysis". *Internet Mathematics*, 7(4):274--298, 2011.

A problem w/ GSEA



- **Its enrichment score considers all genes in C**
- **But …**
  - Not all branches of a large pathway have to "go wrong"
  - ⇒ Cannot detect if only a small part of a pathway malfunctions
- **Solution: Break pathways into subnetworks**

Soh et al. "Finding Consistent Disease Subnetworks Across Microarray Datasets". *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Network-Based Analysis: SNet

- **Group samples into type D and ¬D**
- **Extract & score subnetworks for type D**
  - Get list of genes highly expressed in most D samples
    - **These genes need not be differentially expressed!**
  - Put these genes into pathways
  - Locate connected components (ie., candidate subnetworks) from these pathway graphs
  - Score subnetworks on D samples and on ¬D samples
- **For each subnetwork, compute t-statistic on the two sets of scores**
- **Determine significant subnetworks by permutations**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

SNet: Score Subnetworks

Step 2: Subnetwork Scoring We assign a score vector $SN^{u\_score}_{sn,d}$ with respect to phenotype d to each subnetwork sn within SN^{List} according to Equation 1.

$$SN^{u\_score}_{sn,d} = \langle SN^{i\_score}_{sn,1,d}, SN^{i\_score}_{sn,2,d}, \ldots, SN^{i\_score}_{sn,n,d} \rangle \quad (1)$$

Where n is the number of patients in phenotype d. The formula $SN^{i\_score}_{sn,i,d}$ for the ith patient (also the ith element of this vector) is given by:

$$SN^{i\_score}_{sn,i,d} = \sum_{j=1}^{g} G^{score}_{sn,j,d} \quad (2)$$

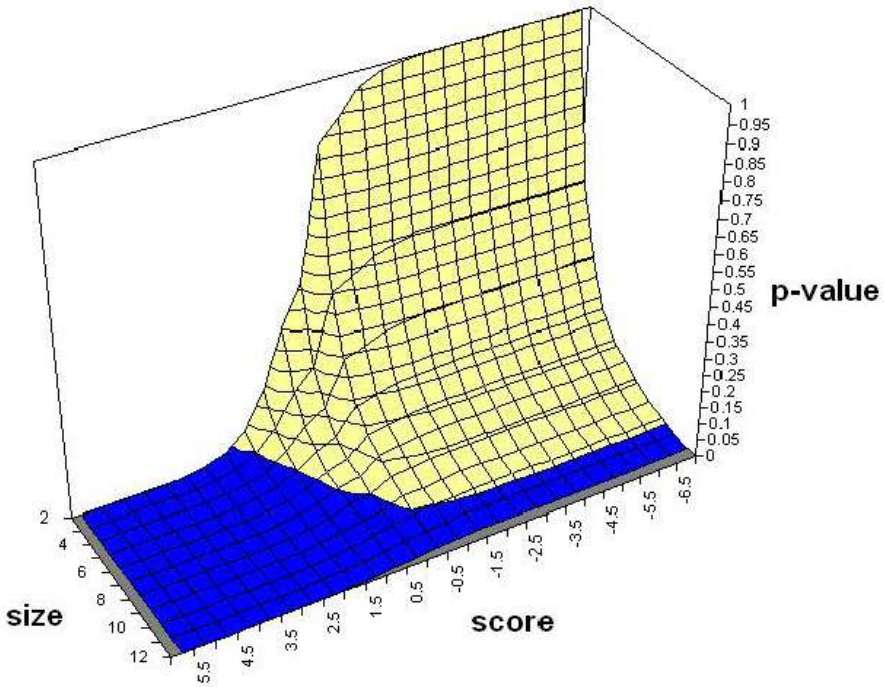
$G^{score}_{sn,j,d}$ refers to the score of the jth gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G^{score}_{sn,j,d}$ is given by Equation 3) and is simply given by:

$$G^{score}_{sn,j,d} = k/n \quad (3)$$

Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d. The entire Step 2 is repeated for the other disease phenotype $\neg d$, giving us the score vectors, $SN^{u\_score}_{sn,d}$ and $SN^{u\_score}_{sn,\neg d}$ for the same set of connected components. The t-test is finally calculated between these two vectors, creating a final t-score for each subnetwork sn within SN_{List}.

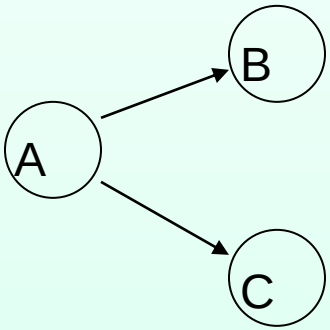
SNet: Significant Subnetworks

- **Randomize patient samples many times**
- **Get t-score for subnetworks from the randomizations**
- **Use these t-scores to establish null distribution**
- **Filter for significant subnetworks from real samples**



Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Key Insight # 1



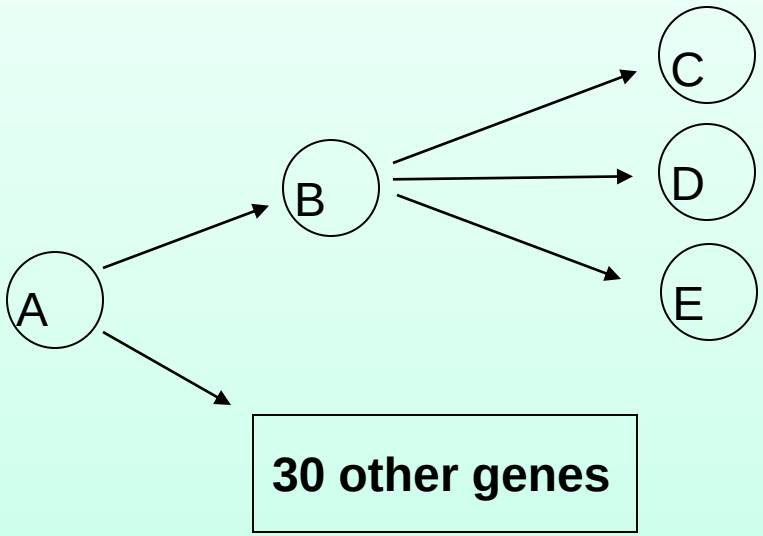
Genes A, B, C are high in phenotype *D*

A is high in phenotype ~*D* but B and C are not

Conventional techniques: Gene B and Gene C are selected. Possible incorrect postulation of mutations in gene B and C

- **SNet does not require all the genes in subnet to be diff expressed**
- **It only requires the subnet as a whole to be diff expressed**
- **Able to capture entire relationship, postulating a mutation in gene A**

Key Insight # 2



A branch within pathway consisting of genes A, B, C, D and E are high in phenotype *D*

Genes C, D and E not high in phenotype ~*D*

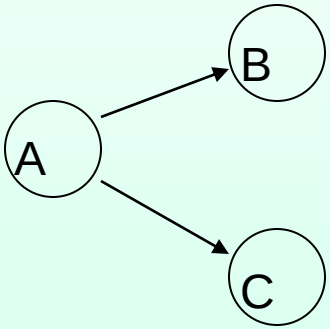
30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

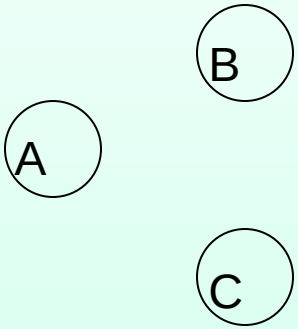
- **SNet: Able to capture the subnetwork branch within the pathway**

Key Insight # 3

Pathway 1



Pathway 2



Genes A, B and C are present in two separate pathways

A, B and C are high in phenotype *D*, but not high in phenotype ~*D*

Conventional techniques:

Both pathways are scored equally. So both got selected, resulting in pathway 2 being a false positive

- **SNet: Able to select only pathway 1, which has the relevant relationship**

Let's see whether SNet gives us subnetworks that are

(i) more consistent between datasets of the same types of disease samples

(ii) larger and more meaningful

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Better Subnetwork Overlap

Table 1. Table showing the percentage overlap significant subnetworks between the datasets. Each row refers to a separate disease (as indicated in the first column). Each disease is tested against two datasets depicted in the second and third column. The overlap percentages refer to the pathway overlaps obtained from running SNet (column 4) and GSEA (column 5) The actual number of overlaps are parenthesized in the same columns.

Disease	Dataset 1	Dataset 2	SNet	GSEA
Leuk	Golub	Armstrong	83.3% (20)	0.0% (0)
Subtype	Ross	Yeoh	47.6% (10)	23.1% (6)
DMD	Haslett	Pescatori	58.3% (7)	55.6% (10)
Lung	Bhatt	Garber	90.9% (9)	0.0% (0)

- **For each disease, take significant subnetworks from one dataset and see if it is also significant in the other dataset**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Better Gene Overlaps

Table 2. Table showing the number and percentage of significant overlapping genes. γ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding algorithms.

Disease	γ	SNet	GSEA	SAM	t-test
Leuk	84	91.3%	2.4%	22.6%	14.3%
Subtype	75	93.0%	4.0%	49.3%	57.3%
DMD	45	69.2%	28.9%	42.2%	20.0%
Lung	65	51.2%	4.0%	24.6%	26.2%

- **For each disease, take significant subnetworks extracted independently from both datasets and see how much their genes overlap**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Larger Subnetworks

Table 3. Table comparing the size of the subnetworks obtained from the t-test and from SNet. The first column shows the disease and the second column shows the number of genes which comprised of the subnetworks. The third and fourth column depicts the number of genes present within each subnetwork for the t-test and SNet respectively. So for instance in the leukemia dataset, we have 8 subnetworks with size 2 genes, 1 subnetwork with size 3 genes for the t-test. For SNet, we have 2 subnetworks with size 5 genes, 3 subnetworks with size 6 genes, 2 subnetworks with size 7 genes and 1 subnetwork with a size of $\geq$ 8 genes

Disease	γ	Num Genes (t-test)				Num Genes (SNet)			
		2	3	4	5	5	6	7	$\geq$ 8
Leuk	84	8	1	0	0	2	3	2	1
Subtype	75	5	1	1	1	1	0	1	6
DMD	45	3	1	0	0	1	0	0	5
Lung	65	3	2	1	0	5	3	0	1

What have we learned?

- **Common headaches in gene expression analysis**
  - Natural fluctuation, protocol noise, batch effect
- **Use of biological background info to tame false positives**
- **Overlap analysis → direct-group analysis → network-based analysis**
- **SNet method yields more consistent and larger disease subnetworks**

References

- Zhang et al. **Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes**. *Bioinformatics*, 25(13):1662-1668, 2009
- [ORA] Khatri & Draghici**. Ontological analysis of gene expression data: Current tools, limitations, and open problems**. *Bioinformatics*, 21(18):3587-3595, 2005
- [FCS] Goeman et al. **A global test for groups of genes: Testing association with a clinical outcome**. *Bioinformatics*, 20(1):93-99, 2004
- [GSEA] Subramanian et al. **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *PNAS*, 102(43):15545-15550, 2005
- [NEA] Sivachenko et al. **Molecular networks in microarray analysis**. *JBCB*, 5(2b):429-546, 2007
- [SNet] Soh et al. **Finding consistent disease subnetworks across microarray datasets**. *BMC Genomics*, **12(Suppl. 13):S15, 2011**

From pathways to models, From static to dynamic:

A couple of very recent papers that are worth your leisure reading…

- Geistlinger et al. **From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems**. *Bioinformatics*, 27(13):i366—i373, 2011
- Zampieri et al. **A system-level approach for deciphering the transcriptional response to prion infection**. *Bioinformatics*, 27(24): 3407--3414, 2011

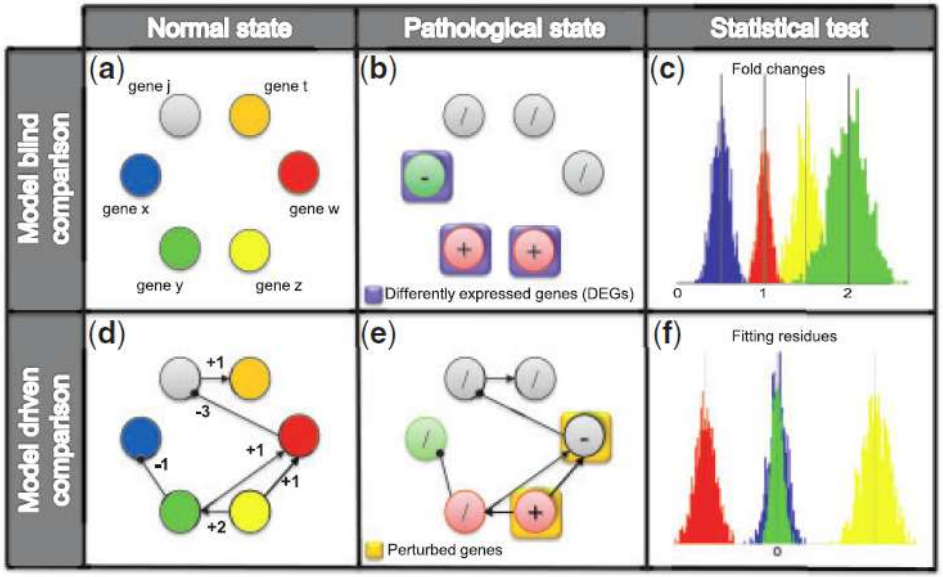


Fig. 1. System response inference: a toy genetic network consisting of six genes exemplifies the advantages of using a system-level data comparison (**a**). Standard statistical tests (i.e. *t*-test) unveil significant fold change in gene expression variations for each transcript individually (**b**), neglecting the underlying regulatory network. Such statistical test can identify whether the expression level of a transcript is significantly changed with respect to a reference. Putative gene expression changes are reported in panel (**c**). In this specific example, two genes are identified to be overexpressed [red/+ nodes] and one downregulated (green/- node), while the remaining three do not show any changes (grey nodes). By knowing the corresponding genetic regulatory network (**d**), we can discriminate the coherent variations from the unexpected ones. As shown in the example, two of the genes that showed a significant expression variations are consistent with model predictions i.e. the expression changes of genes *x* and *y* can be explained by the variation of gene *z*. This is reflected by a skew distribution of discrepancies (i.e. residues), between model predictions and observed data, centered around 0 (**f**). At the same time, one transcript, *w*, is not responding coherently to the initial model. The fact that its expression is unchanged, when it should have been increased, might relate to an anomalous direct effect of the pathology, preventing a synergistic response between all the genes in the system. Hence, the list of 'perturbed genes' can be sensibly different from the standard DEGs identified from individual fold change analysis (**b**/**e**).

Using Biological Networks, Part 2: *Delivering More Powerful Proteomic Profile Analysis*

Limsoon Wong

NUS
National University of Singapore

- **First, some basics of proteomic MS…**

Typical Proteomic MS Experiment

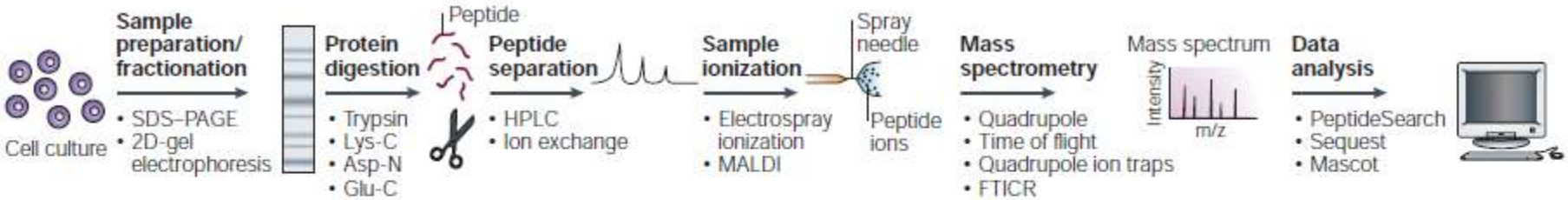


Figure 1 | **The mass-spectrometry/proteomic experiment.** A protein population is prepared from a biological source — for example, a cell culture — and the last step in protein purification is often SDS–PAGE. The gel lane that is obtained is cut into several slices, which are then in-gel digested. Numerous different enzymes and/or chemicals are available for this step. The generated peptide mixture is separated on- or off-line using single or multiple dimensions of peptide separation. Peptides are then ionized by electrospray ionization (depicted) or matrix-assisted laser desorption/ionization (MALDI) and can be analysed by various different mass spectrometers. Finally, the peptide-sequencing data that are obtained from the mass spectra are searched against protein databases using one of a number of database-searching programmes. Examples of the reagents or techniques that can be used at each step of this type of experiment are shown beneath each arrow. 2D, two-dimensional; FTICR, Fourier-transform ion cyclotron resonance; HPLC, high-performance liquid chromatography.

Source: Steen & Mann. The ABC's and XYZ's of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

Diagnosis Using Proteomics

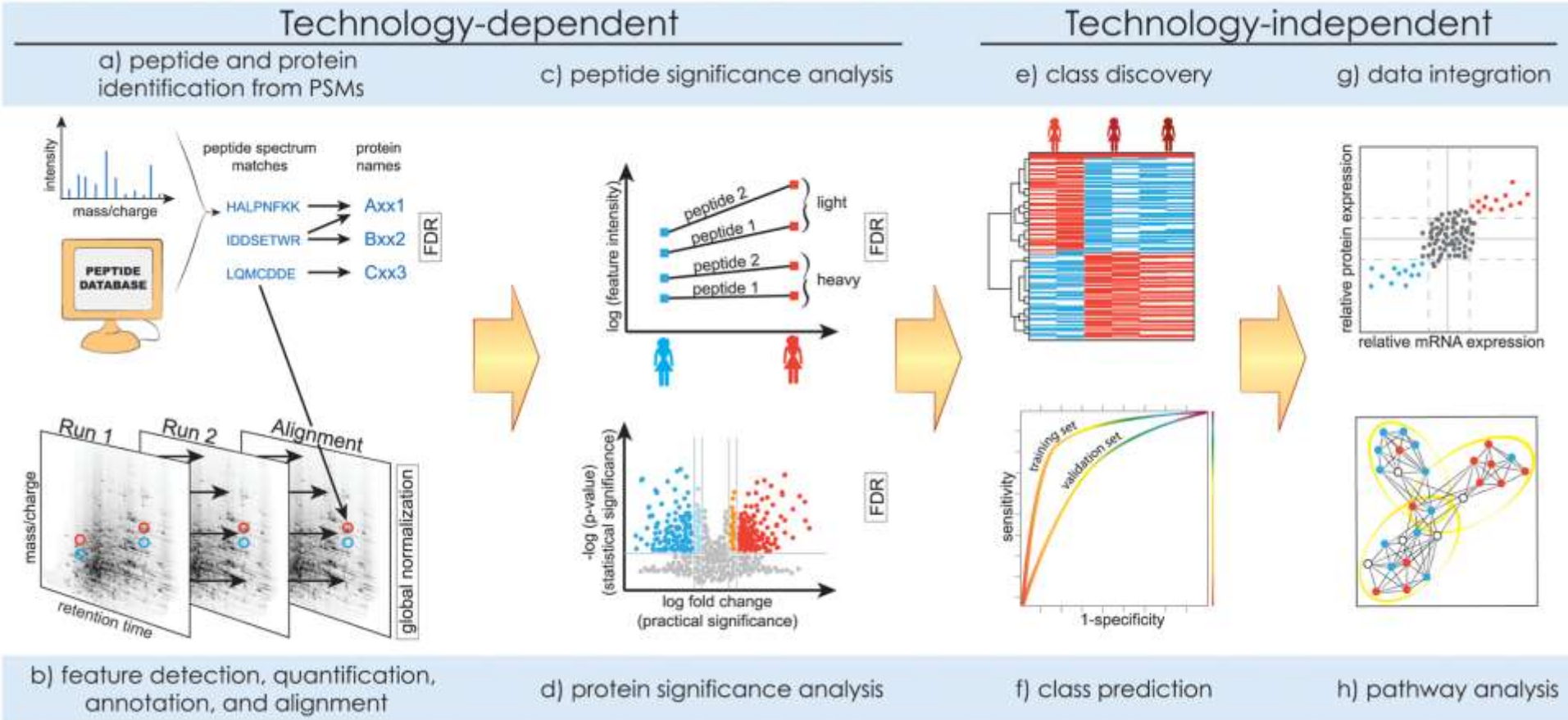


Image credit: Kall and Vitek, *PLoS Comput Biol* , 7(12): e1002277, 2011

A rather nice set of proteomic profiles of leukemia patients

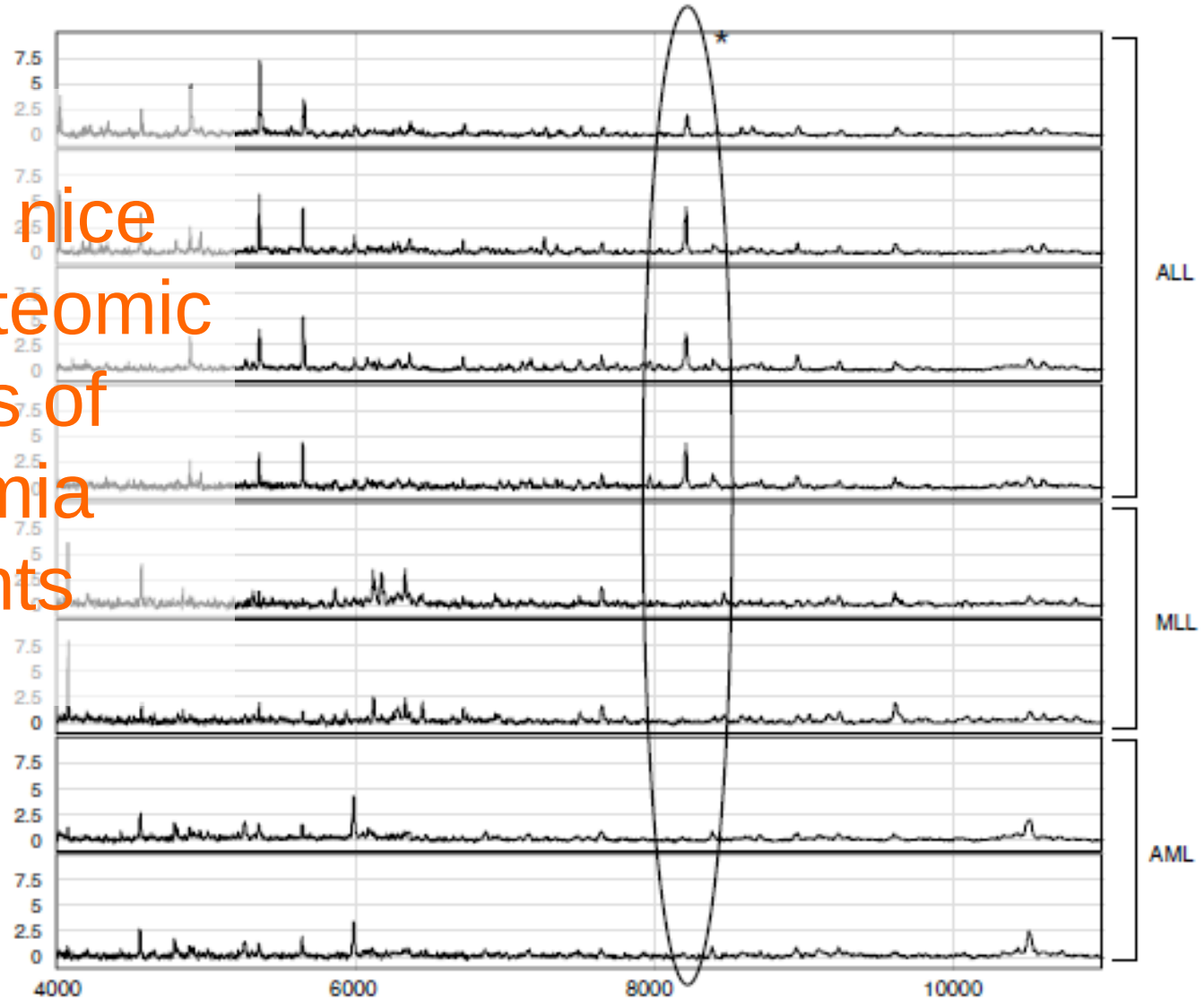


Figure 1 Spectra from SELDI-TOF MS analysis of REH, 697, MV4;11, and Kasumi cell lines. Protein (4 μg) from each cell type was analyzed on SAX2 ProteinChip® Arrays. ALL cell lines shown are REH and 697, the MLL cell line is MV4;11, and the AML cell line is Kasumi. The asterisk indicates the differentially expressed protein at 8.3 kDa.

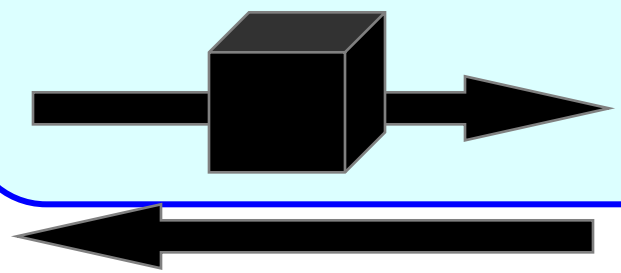
Source: Hegedus et al. Proteomic analysis of childhood leukemia. Leukemia, 19:1713-1718, 2005

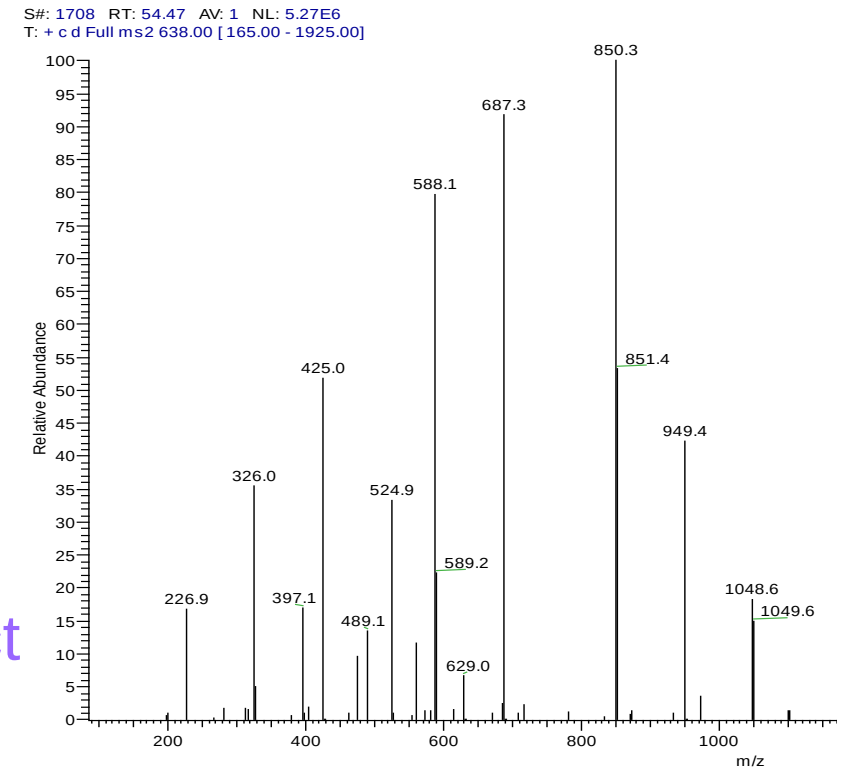
Protein Identification by Mass Spec

Sequence

Step 1:

MS/MS instrument





Database search

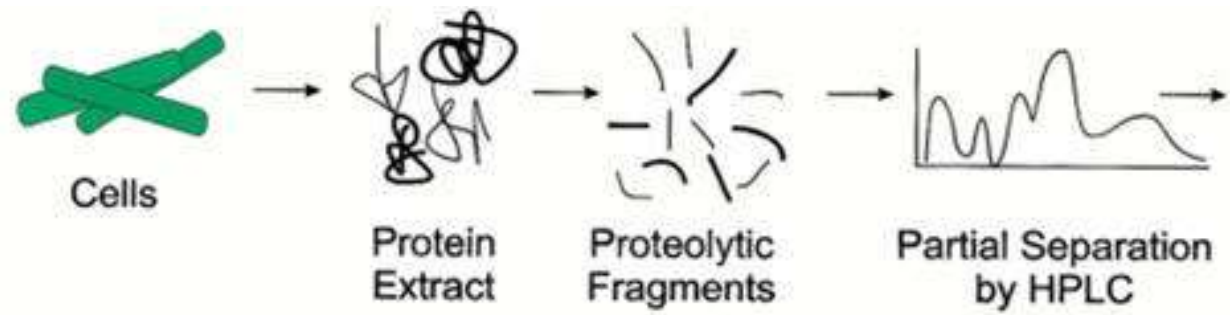
- Sequest, Mascot, InSpect

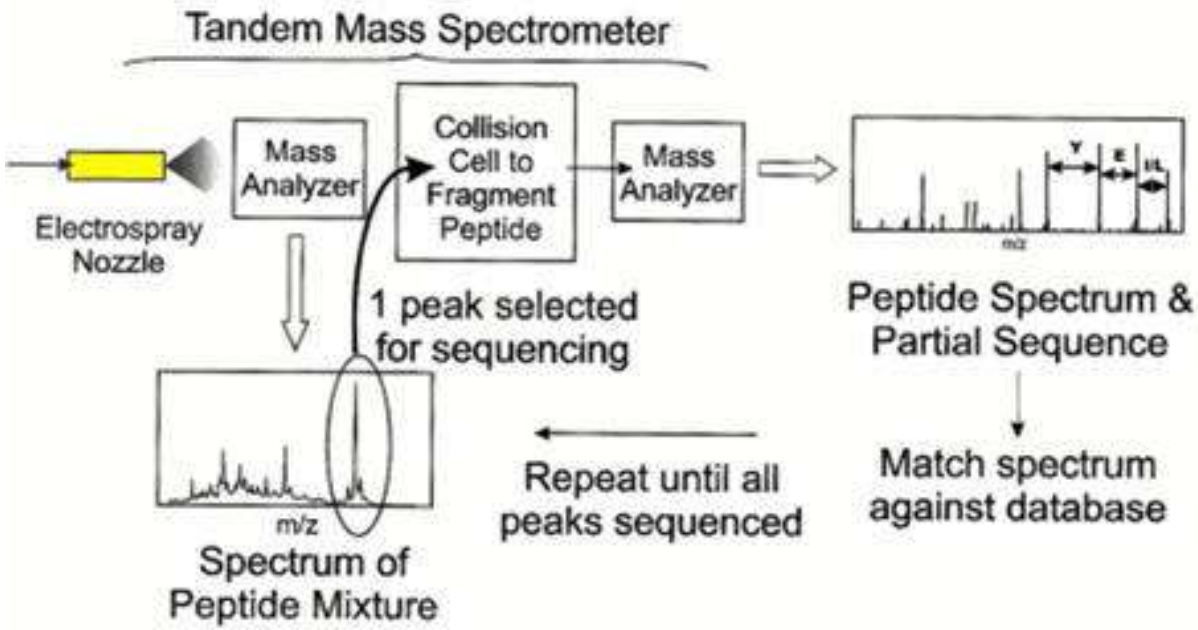
de Novo interpretation

- Lutefisk, Peaks, PepNovo

Source: Leong Hon Wai

Tandem Mass-Spectrometry





Source: Leong Hon Wai

Breaking Protein into Peptides, and Peptides into Fragment Ions

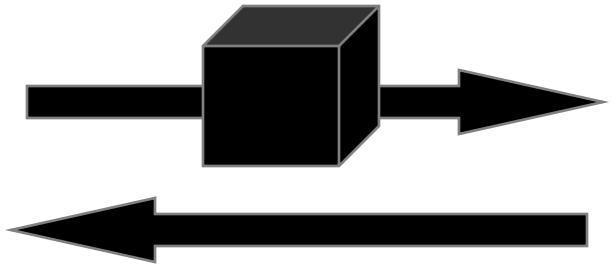
- **Proteases, e.g. trypsin, break protein into peptides**
- **A Tandem Mass Spectrometer further breaks the peptides down into fragment ions and measures the mass of each piece**
- **Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones**
- **Mass Spectrometer measures mass/charge ratio of an ion**

Source: Leong Hon Wai

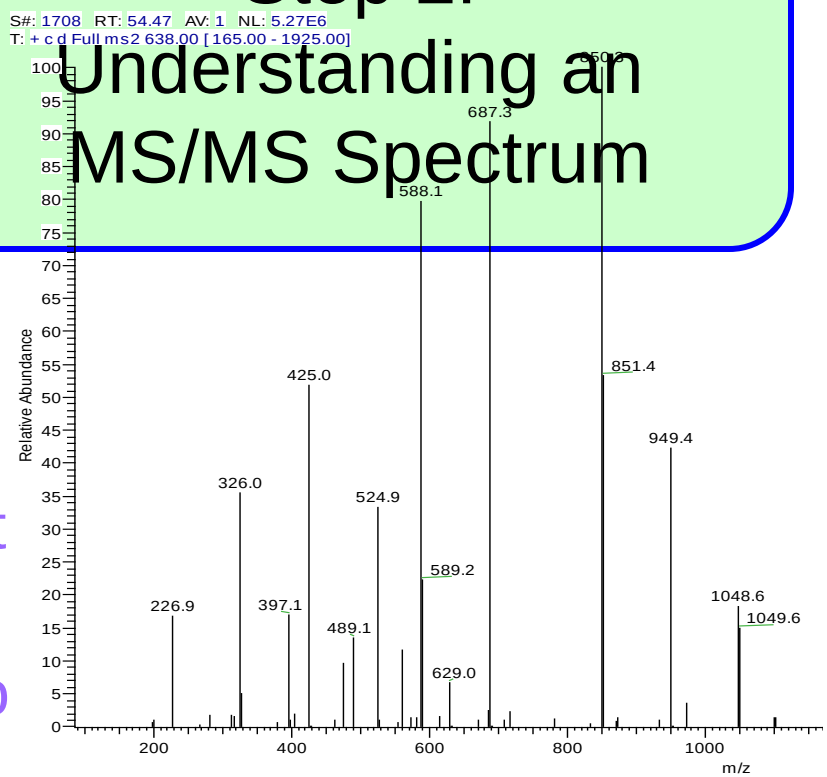
Peptide Identification by Mass Spec

Sequence

MS/MS instrument



Step 2: Understanding an MS/MS Spectrum



Database search

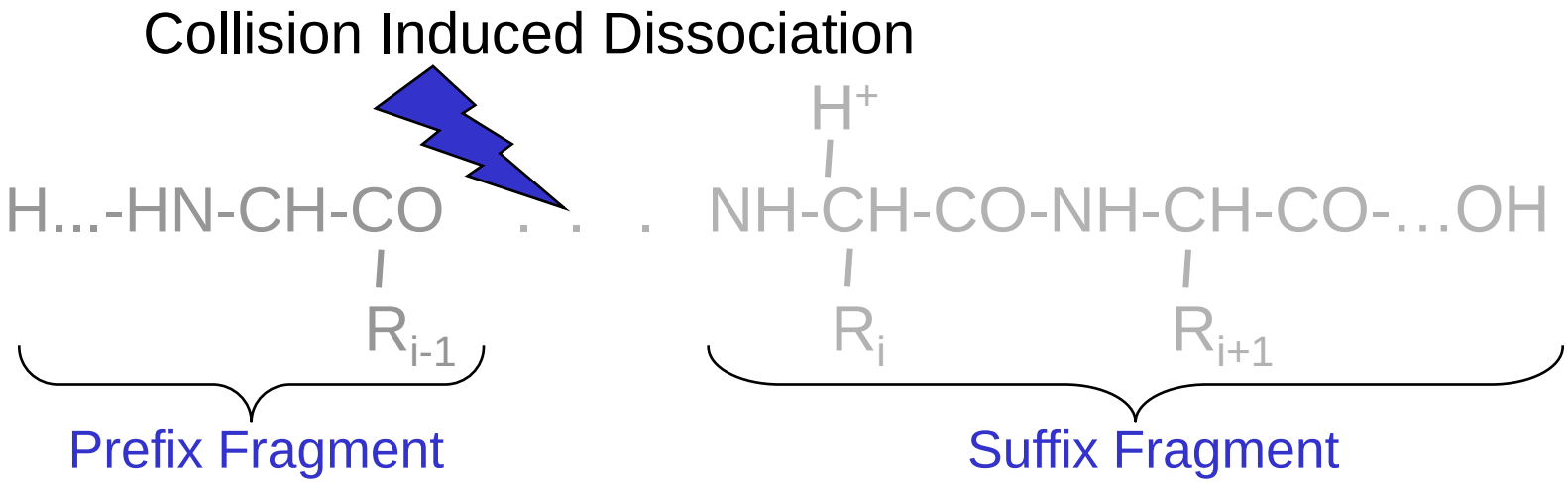
- Sequest, Mascot, InSpect

de Novo interpretation

- Lutefisk, Peaks, PepNovo

Source: Leong Hon Wai

Peptide Fragmentation



- **Peptides tend to fragment along the backbone**
- **Fragments can also loose neutral chemical groups like NH_3 and H_2O**

Source: Leong Hon Wai

Bafna & Edwards. “On de novo interpretation of tandem mass spectra for peptide identification”. RECOMB 2003, pp. 9-18

Peptide Fragmentation

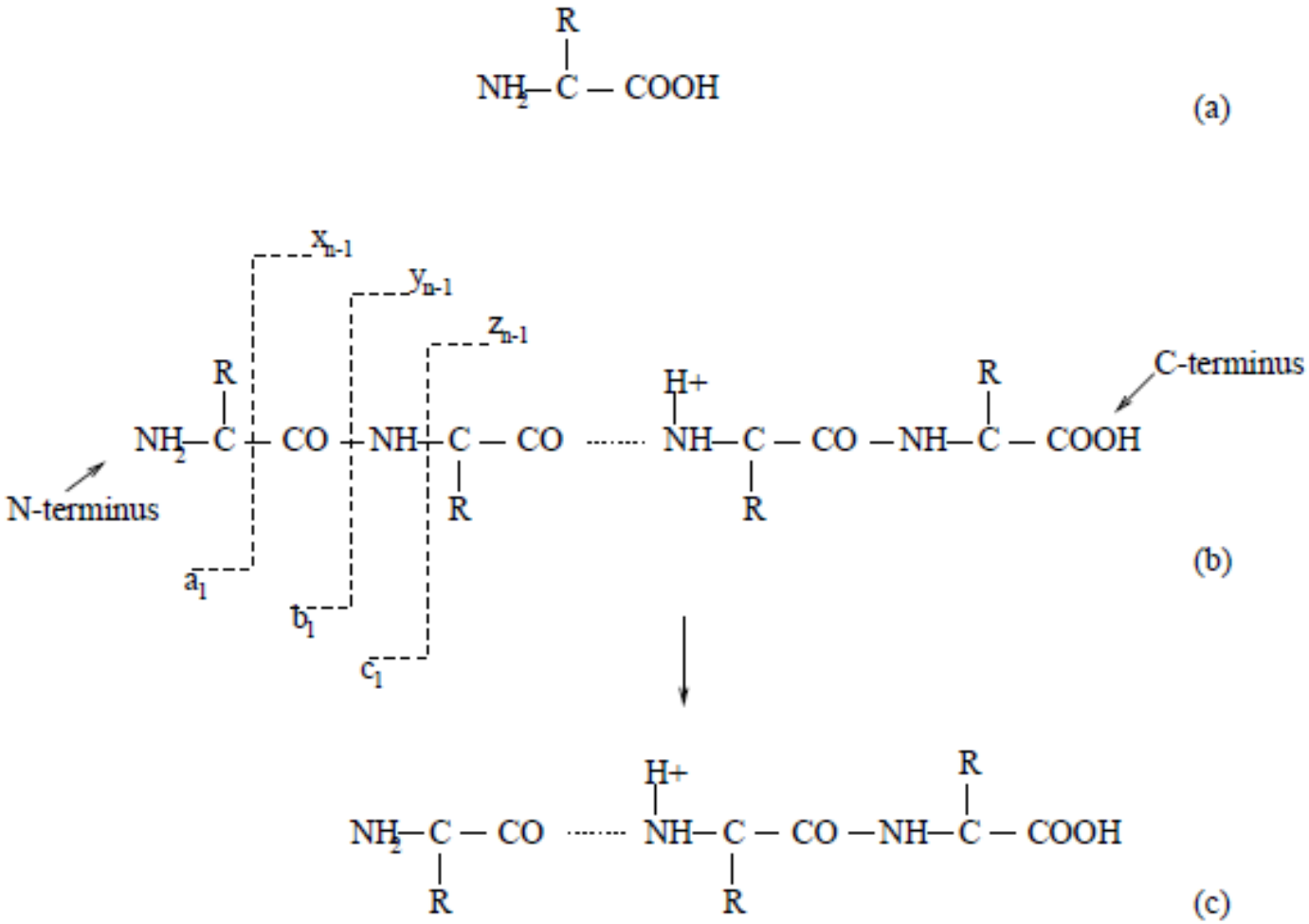
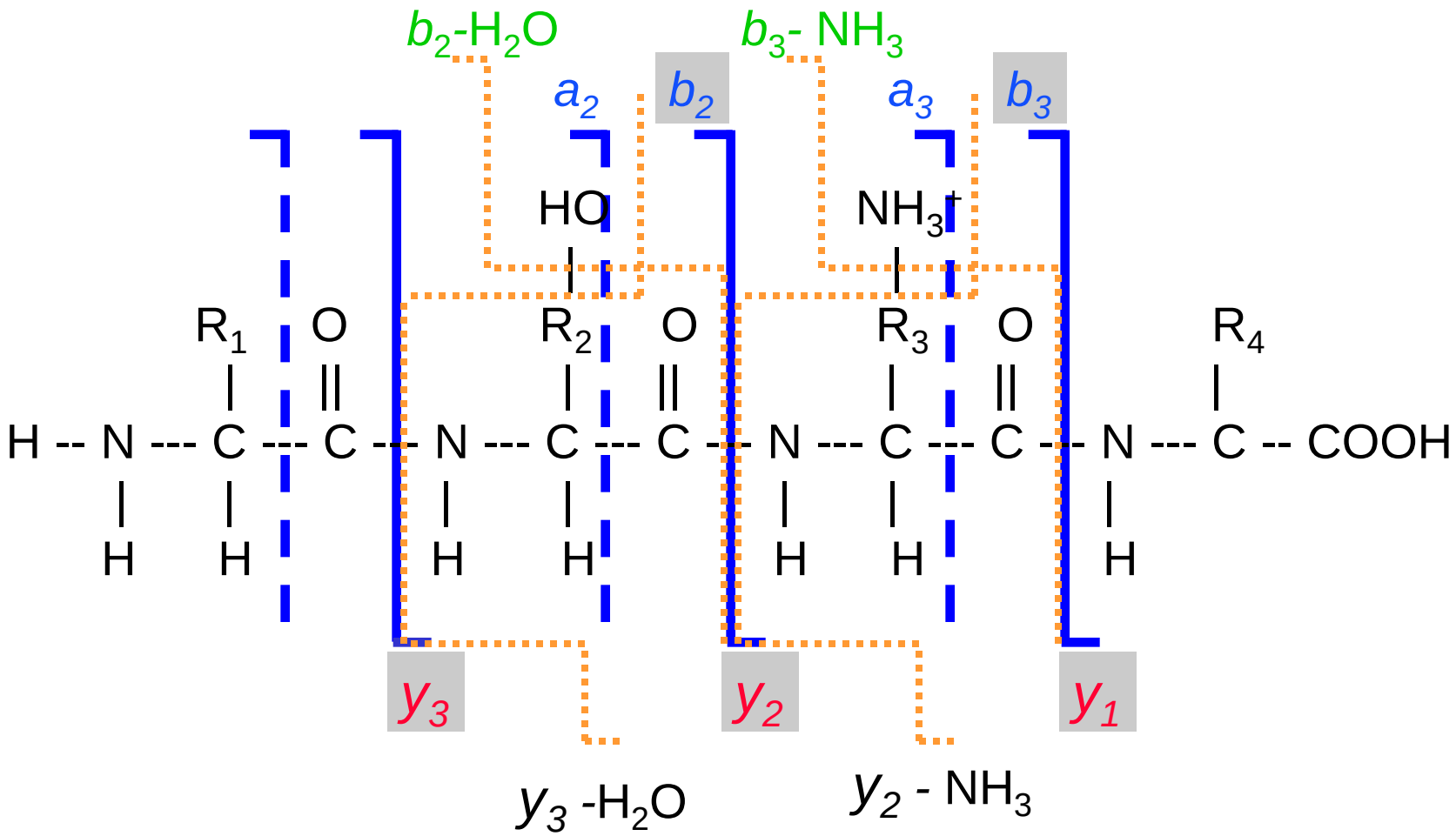


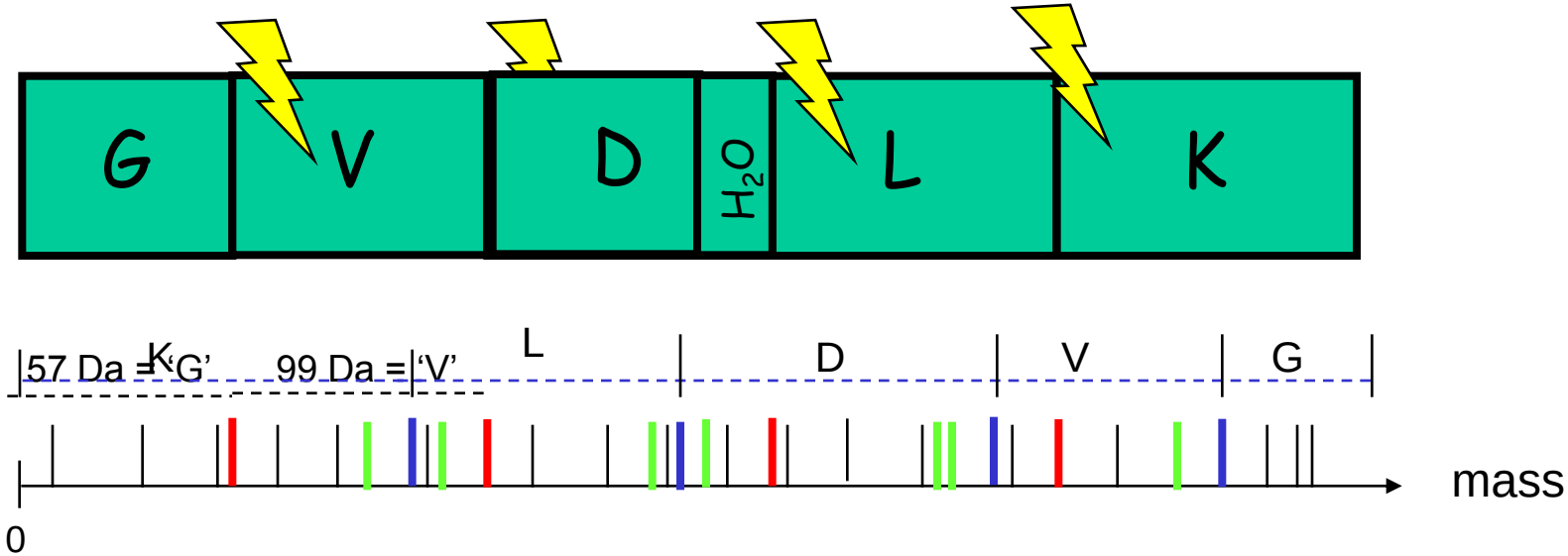
Figure 1: (a) The structure of an amino-acid. (b) An ionized peptide. (c) y^+_{n-1} ion

… and fragments due to neutral losses



Source: Leong Hon Wai

Mass Spectra



- **The peaks in the mass spectrum:**
  - Prefix and Suffix Fragments
  - Fragments with neutral losses ($-H_2O$, $-NH_3$)
  - Noise and missing peaks

Source: Leong Hon Wai

Bafna & Edwards. “On de novo interpretation of tandem mass spectra for peptide identification”. RECOMB 2003, pp. 9-18

Example MS/MS Spectrum

88	145	292	405	534	663	778	924	b-ions
S	G	F	L	E	E	D	K	
924	837	780	633	520	391	262	141	y-ions

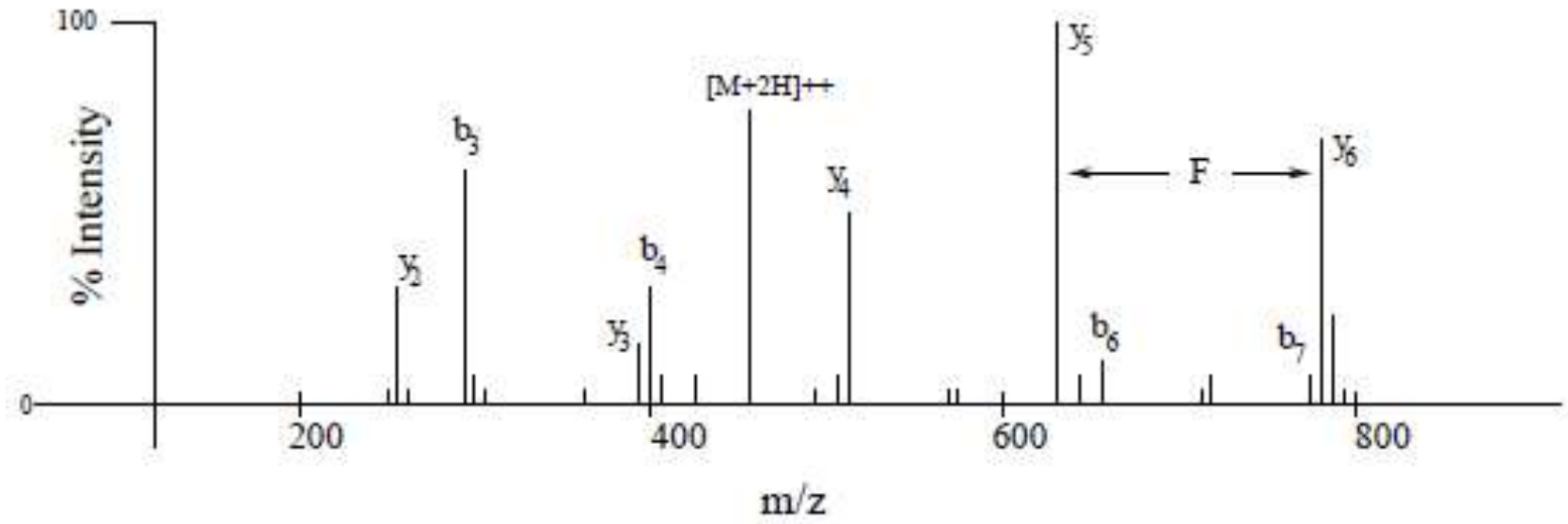
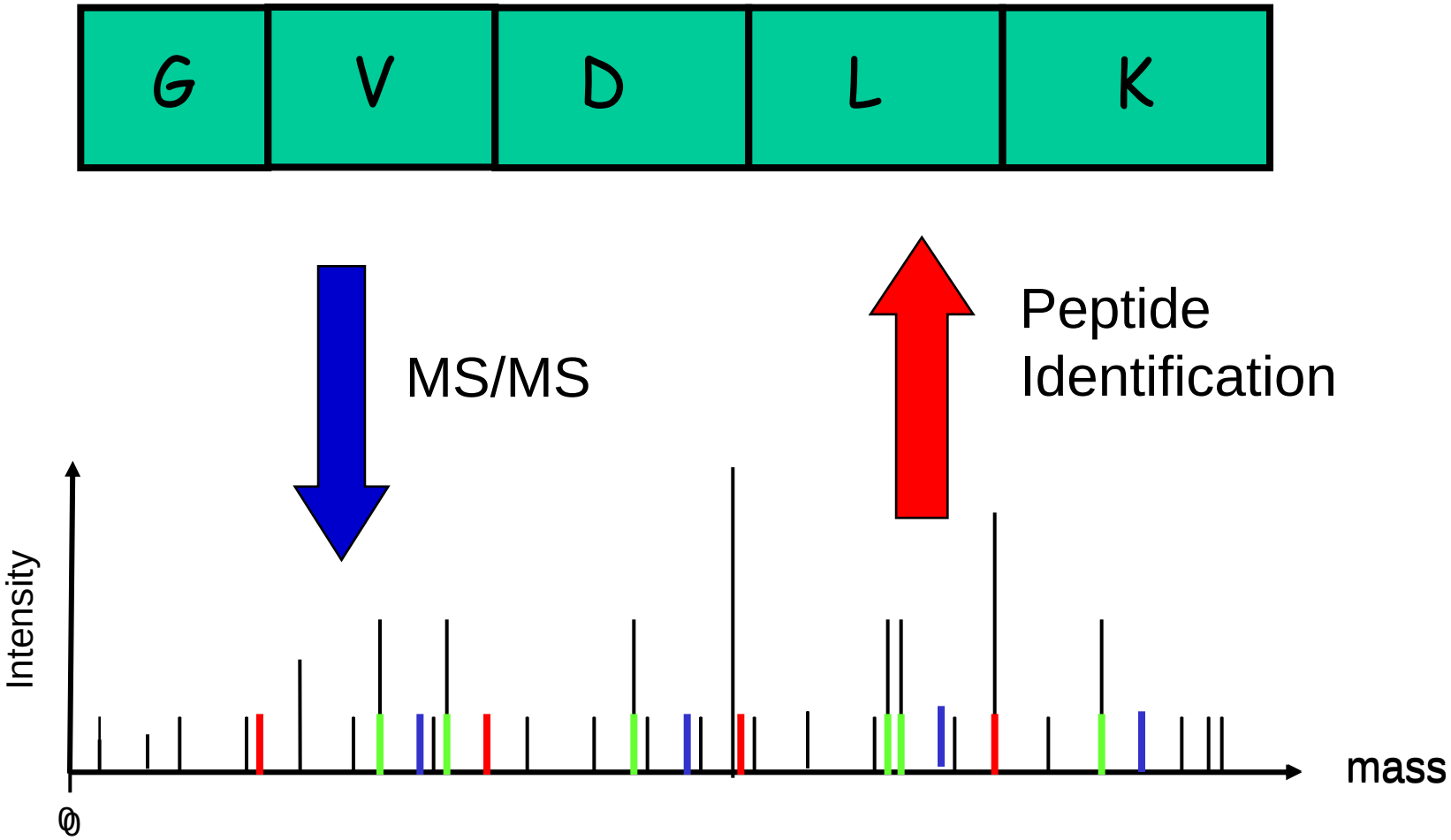


Figure 2: MS/MS spectrum for peptide SGFLEEDK.

Protein Identification with MS/MS

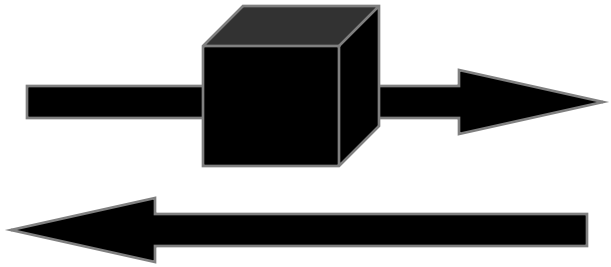


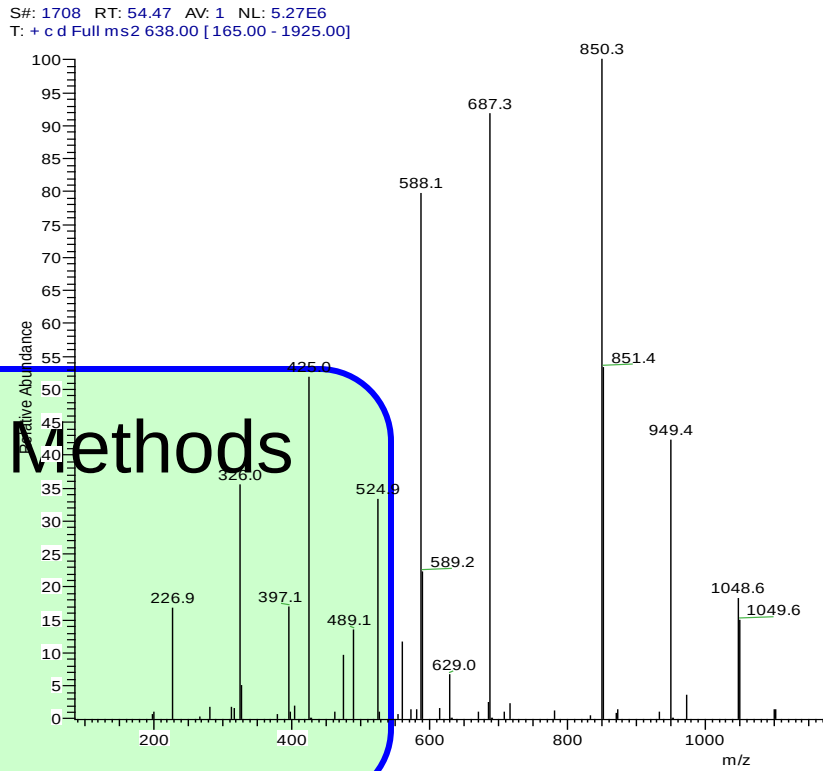
Source: Leong Hon Wai

Peptide Identification by Mass

Sequence

MS/MS instrument





Step 3: Computational Methods

Database search

Sequest, Mascot

de Novo interpretation

Lutefisk, Peaks, PepNovo

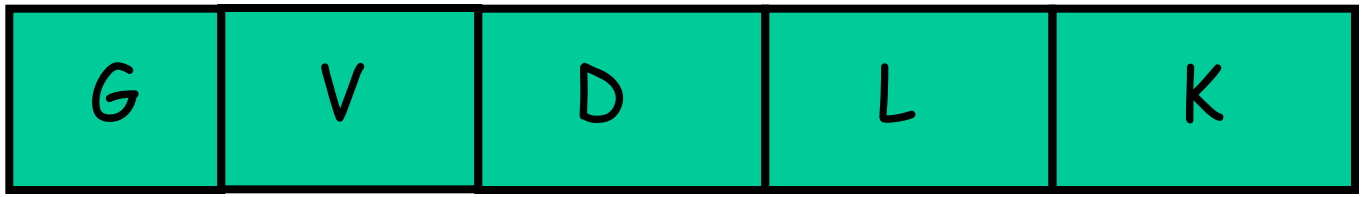
Source: Leong Hon Wai

Database Search Algorithms

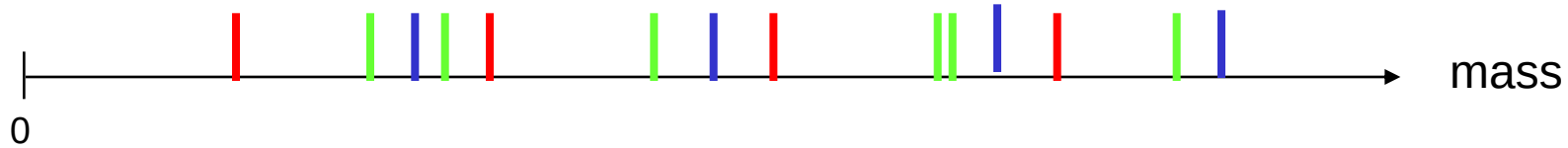
- **Database search**
  - Used for spectrum from known peptides
  - Rely on completeness of database
- **General Approach**
  - Match given spectrum with known peptide
  - Enhanced with advanced statistical analysis and complex scoring functions
- **Methods**
  - SEQUEST, MASCOT, InsPecT, Paragon

Theoretical Spectrum for a Peptide

- **Given this peptide**



- **Its theoretical spectrum is**

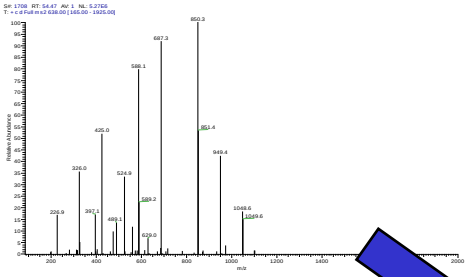


- **Theoretical spectrum is dependent on**
  - Set of ion-types considered
  - Larger if multi-charge ions are considered

Source: Leong Hon Wai

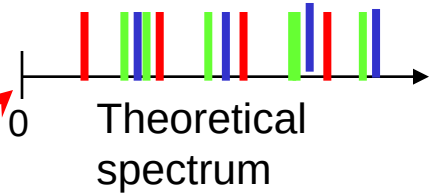
Database Search Algorithm

Database Search



Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIIMNVRT, AVGELTK, HEWAILF, GHNLWAMNAC, GVFGSVLRA, EKLNKAATYIN..



Match

Matching Score for this peptide

Repeat for all the peptides in the Database

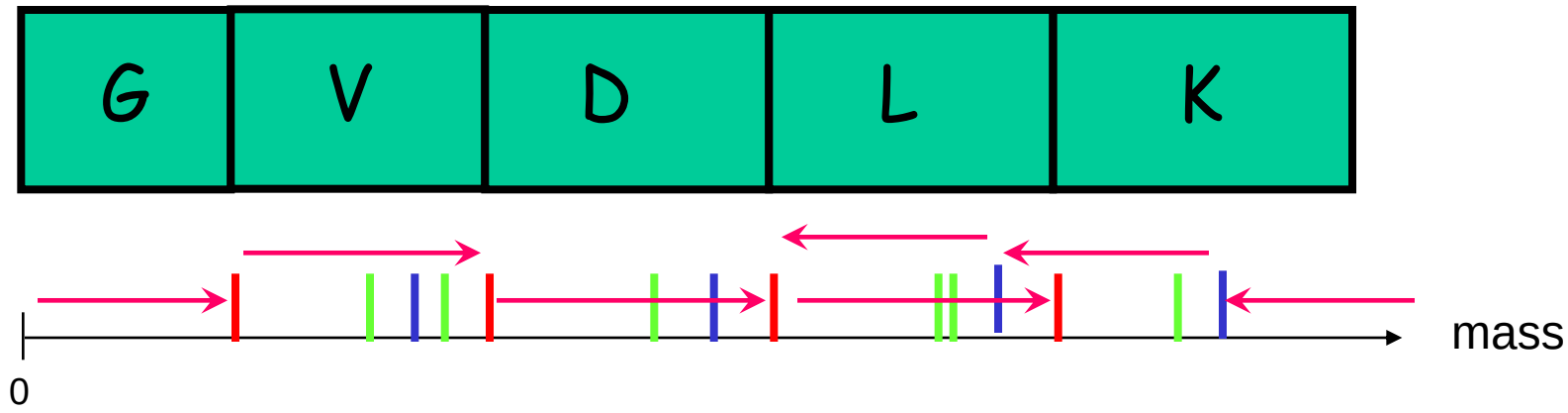
Source: Leong Hon Wai

De Novo Sequencing Algorithms

- **Given a spectrum**
  - Build a spectrum graph
  - Peptides are paths in this graph
  - Find the best path

Source: Leong Hon Wai

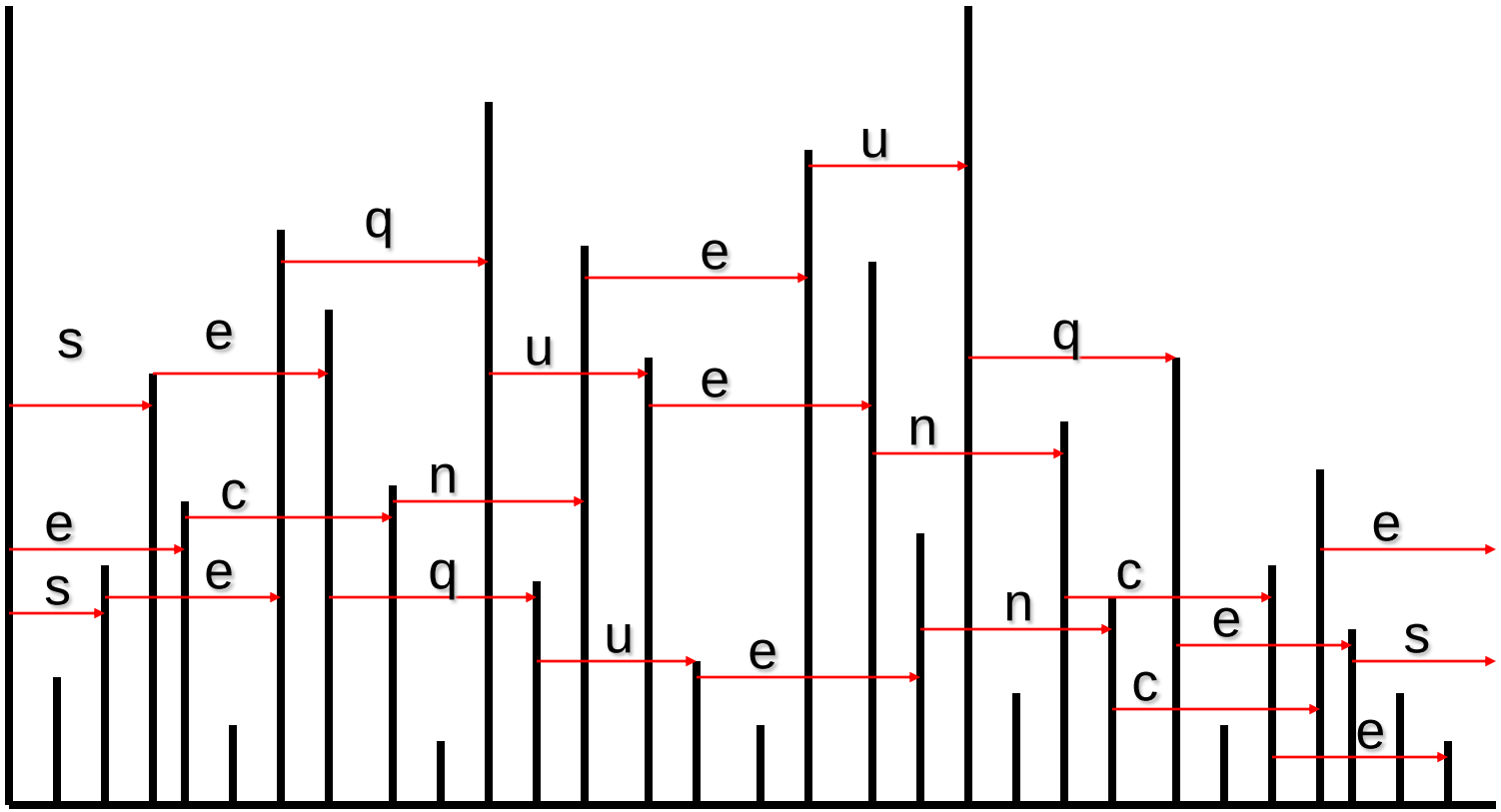
Spectrum Graph for a Peptide



- **Connect peaks together**
  - If their mass difference = mass of an amino acid
- **Theoretical spectrum is dependent on**
  - Set of ion-types considered
  - Larger if multi-charge ions are considered

Source: Leong Hon Wai

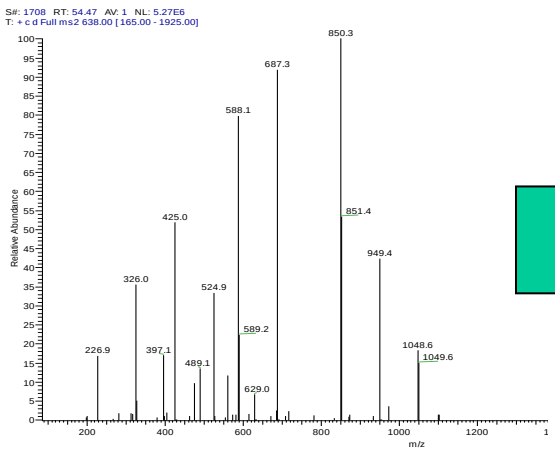
Building a Graph from a Spectrum

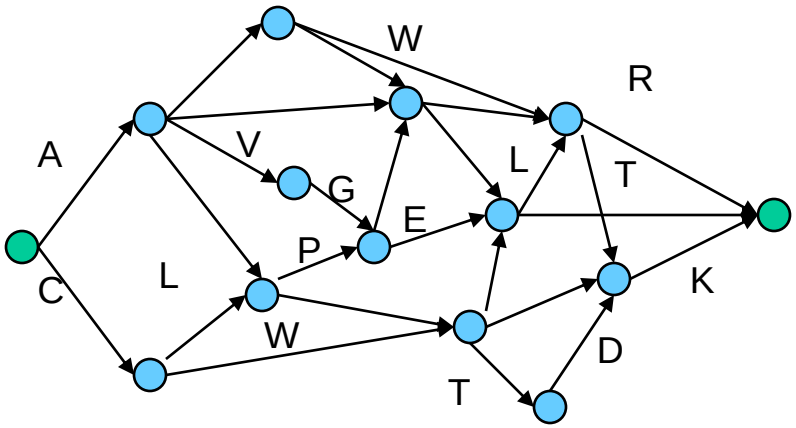


Source: Leong Hon Wai

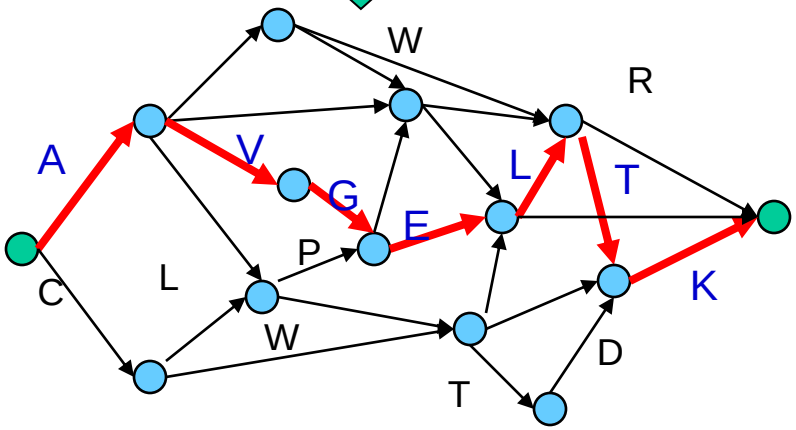
Frank, et al. "De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry". J. Proteome Res. 6:114-123, 2007

De Novo Sequencing Algorithms





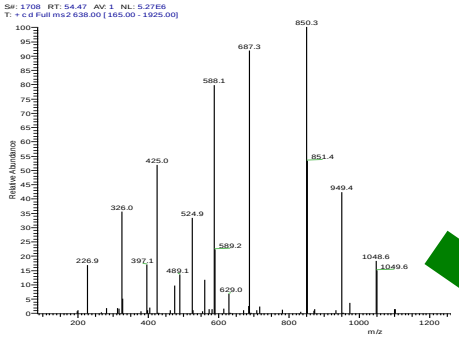
Find longest directed acyclic path



AVGELTK

De Novo vs. Database Search

Database Search

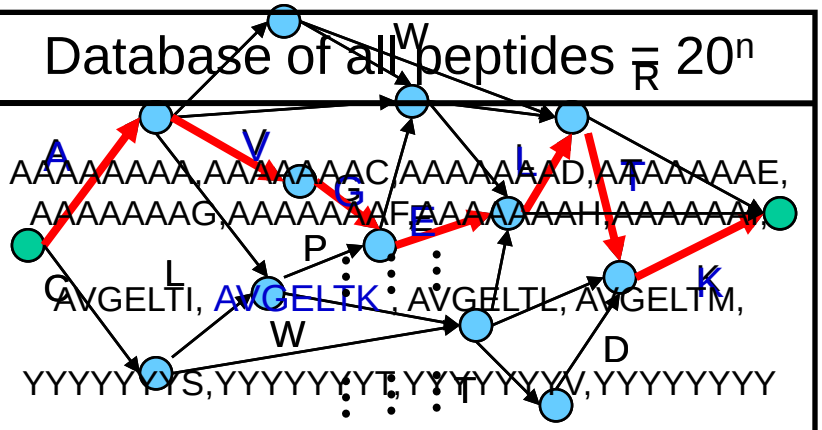


De Novo

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIIMNVRT, AVGELTK, HEWAILF, GHNLWAMNAC, GVFGSVLRA, EKLNKAATYIN..

Database of all peptides = 20^n



AVGELTK

Source: Leong Hon Wai

De Novo vs. Database Search: A Paradox

- **The database of all peptides is huge ≈ $O(20^n)$**
- **The database of all known peptides is much smaller ≈ $O(10^8)$**

- **However, de novo algorithms can be much faster, even though their search space is much larger!**
  - A database search scans all peptides in the search space to find best one
  - De novo eliminates the need to scan all peptides by modeling the problem as a graph search

Source: Leong Hon Wai

Protein Identification

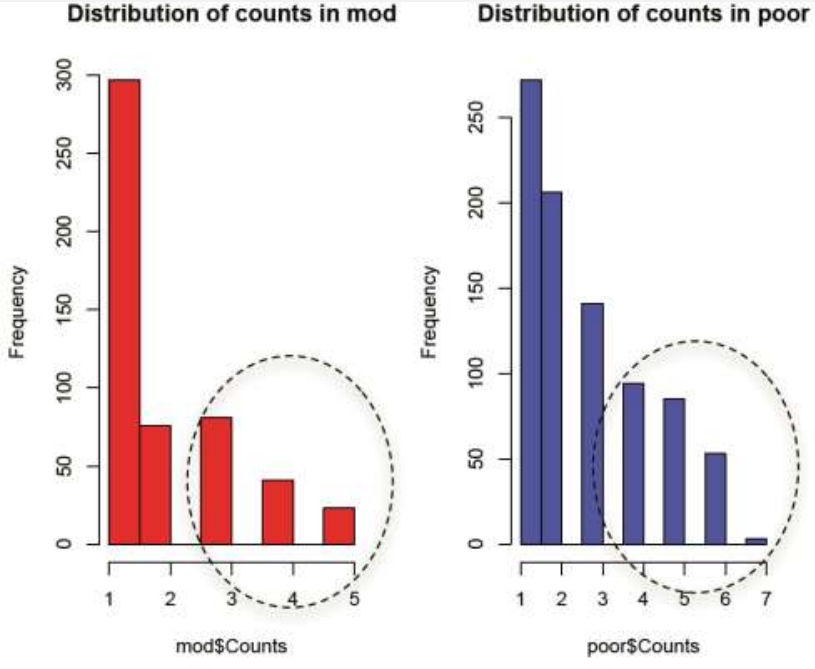
- **After all the peptides have been identified, they are grouped into protein identifications**
- **Peptide scores are added up to yield protein scores**
- **Confidence of a particular peptide identification increases if other peptides identify the same protein and decreases if no other peptides do so**
- **Protein identifications based on single peptides should only be allowed in exceptional cases**

Source: Steen & Mann. The ABC's and XYZ's of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

Cf. Gene Expression Profile Analysis

- **Once the proteins are identified, the proteomic profile of a sample can be constructed**
  - I.e., which protein is found in the sample and how abundant it is
- **Similar to gene expression profile. So gene expression profile analysis techs can be applied**
- **Some key differences**
  - Proteomic profile has much fewer features
  - Proteomic profiling study has much fewer samples

Part 2: Delivering more powerful proteomic profile analysis



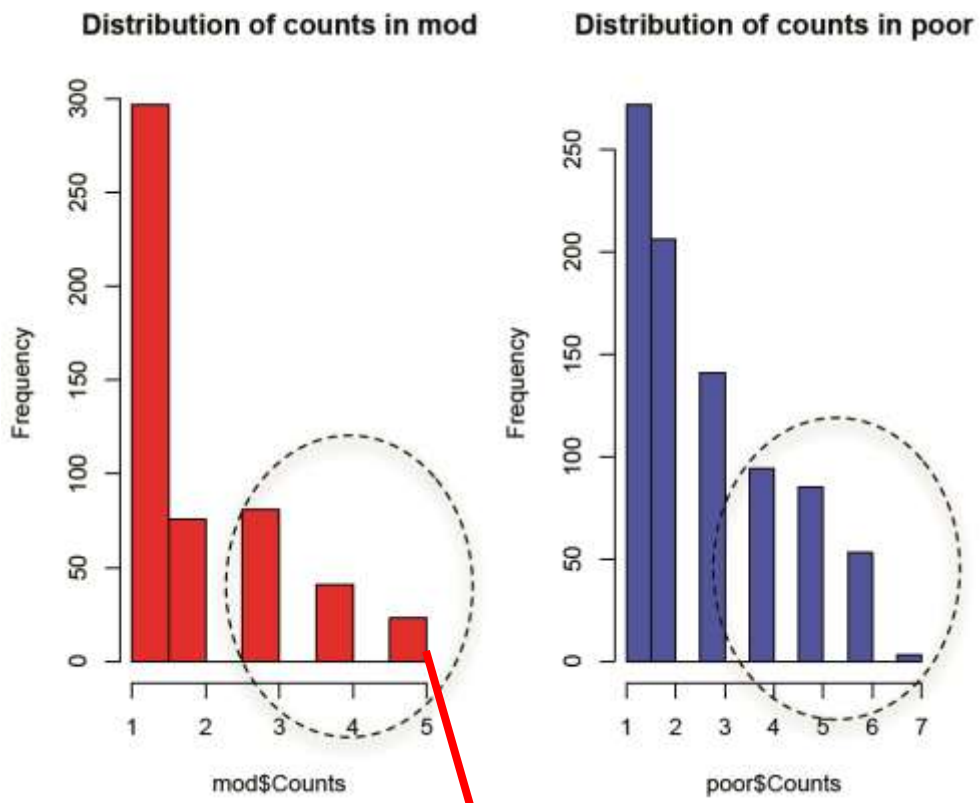
- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP
  - PDS
- **Improving coverage**
  - CEA
  - PEP
  - Max Link

Peptide & protein identification by MS is still far from perfect

- **"… peptides with low scores are, nevertheless, often correct, so manual validation of such hits can often 'rescue' the identification of important proteins."**

Steen & Mann. **The ABC's and XYZ's of peptide sequencing**. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

Typical frequency distribution of proteins detected in proteomic profiles



Only 25 out of 800+ proteins are common to all 5 mod-stage HCC patients!

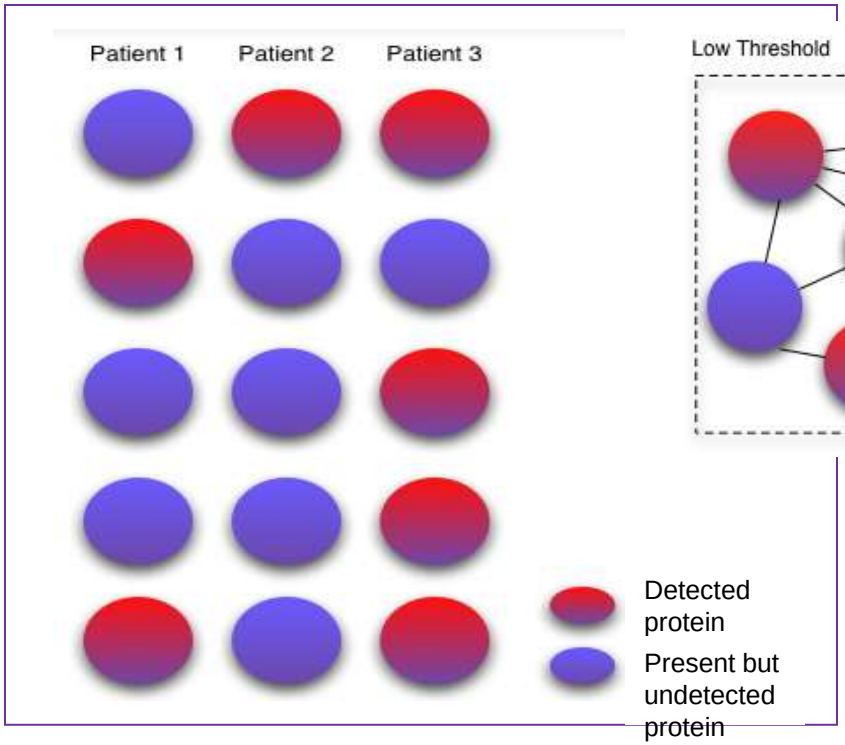
Image credit: Wilson Goh

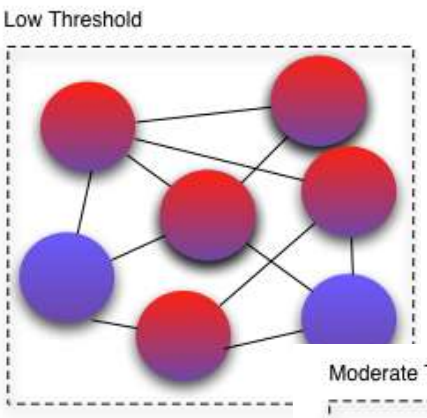
Issues in Proteomic Profiling

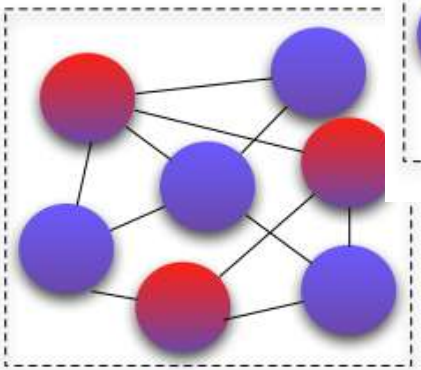
- **Coverage**
- **Consistency**

⇒ **Thresholding**

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







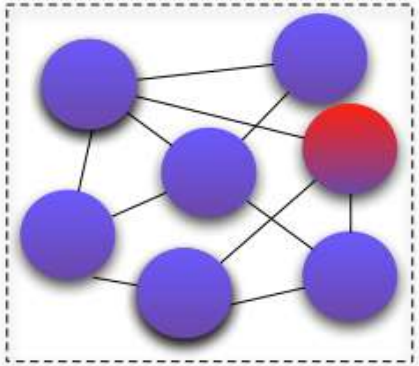
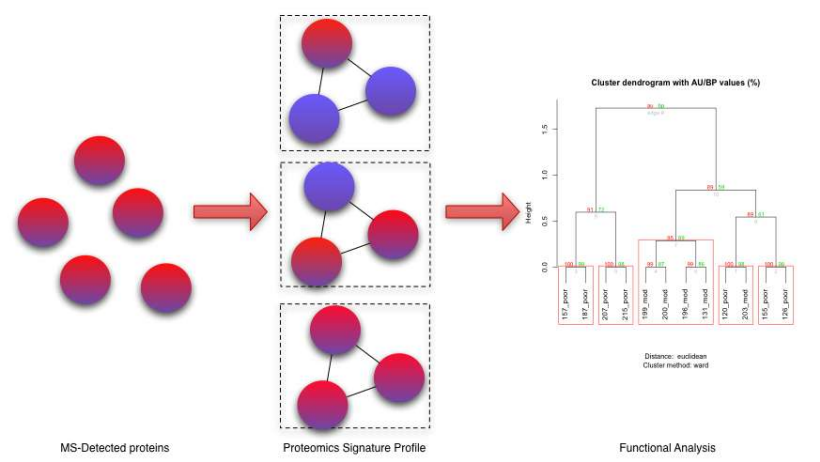


Image credit: Wilson Goh

Part 2: Delivering more powerful proteomic profile analysis

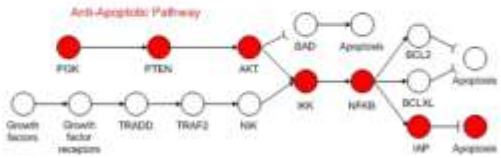


- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP
  - PDS
- **Improving coverage**
  - CEA
  - PEP
  - Max Link

An inspiration from gene expression profile analysis

11

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

Co… 2011 © Limsoon W

12

Taming false positives by considering pathways instead of all possible groups

Group of Genes

- Suppose
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- What is the chance of a group of 5 genes being perfectly correlated to these samples?

- Prob(group of genes correlated) = $(1/2^6)^5$
  - Good, << $1/2^6$
- ~~# of groups = $^{100000}C_5$~~
- ~~E(# of groups of genes correlated) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$~~

# of pathways = 1000

E(# of pathways correlated) = $1000 * (1/2^6)^5 = 9.3*10^{-7}$

⇒ Even more false positives?

- Perhaps no need to consider every group

Microarray Workshop for Gene Expression Profiling, NUH, 23/9/2011 Copyright 2011 © Limsoon Wong

Copyright 2011 © Limsoon Wong

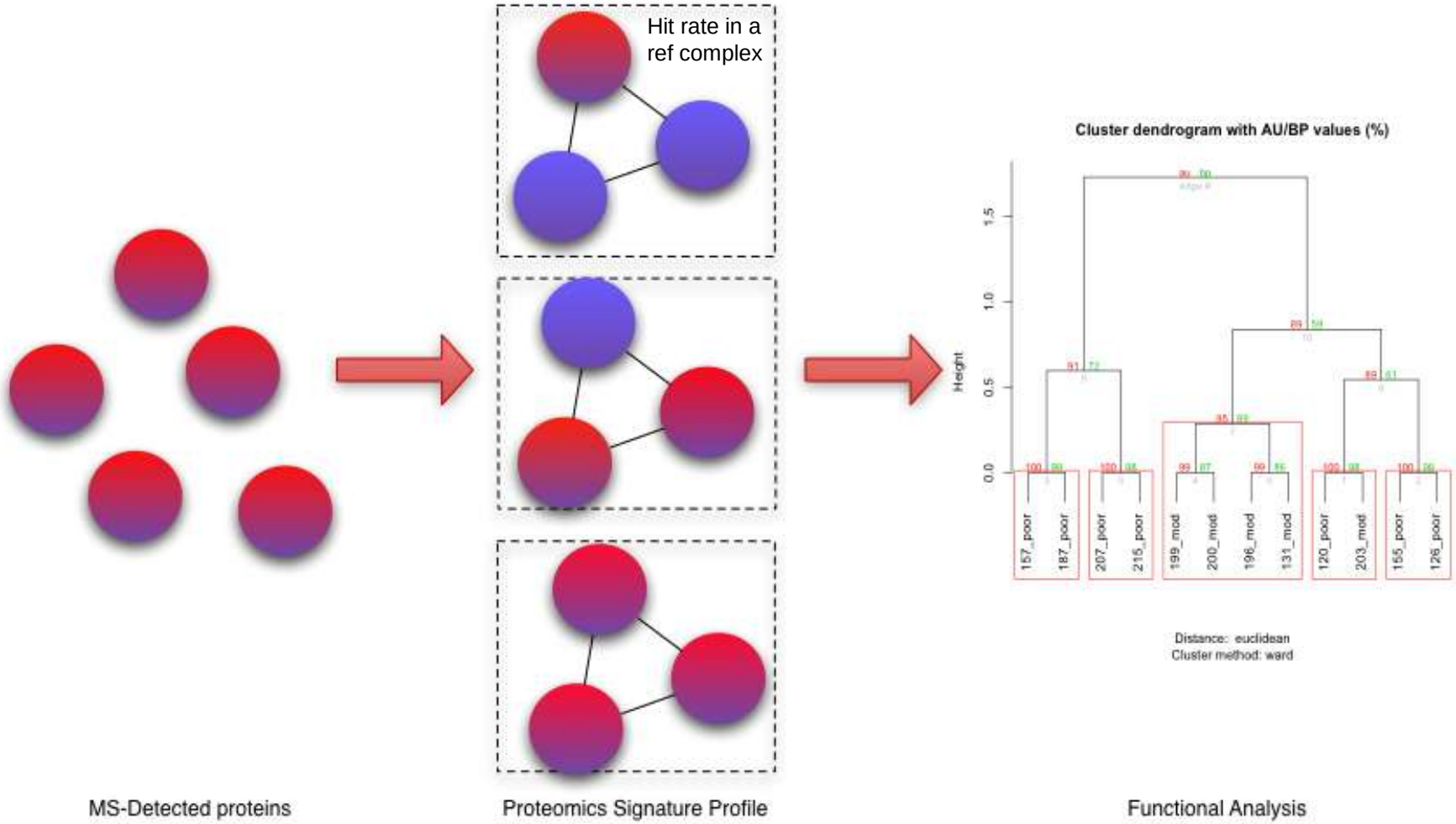
Contextualization!

We try an adaptation of SNet on proteomics profiles…

“Proteomic Signature Profiling” (PSP)

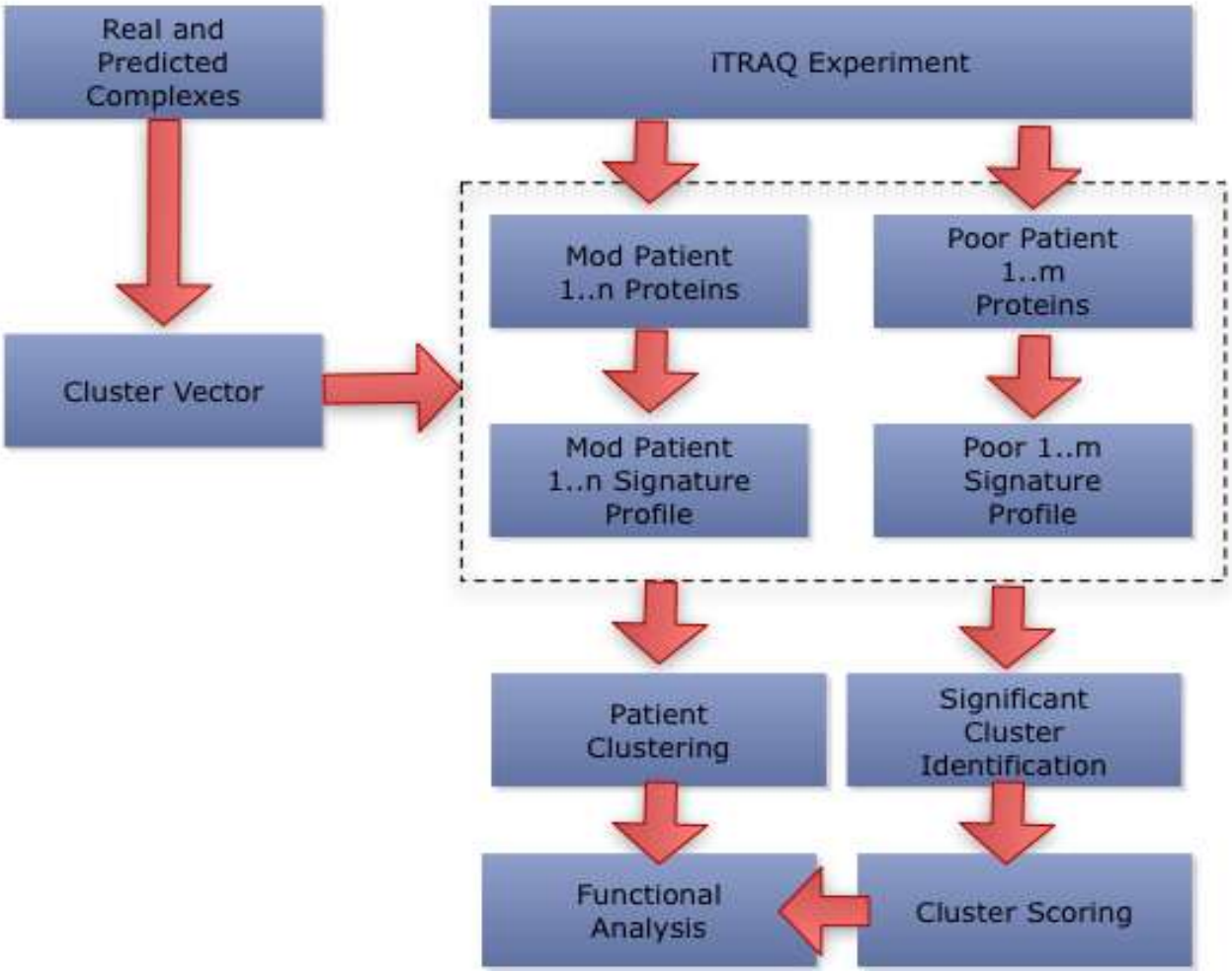
Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 1st revision.

"Threshold-free" Principle of PSP



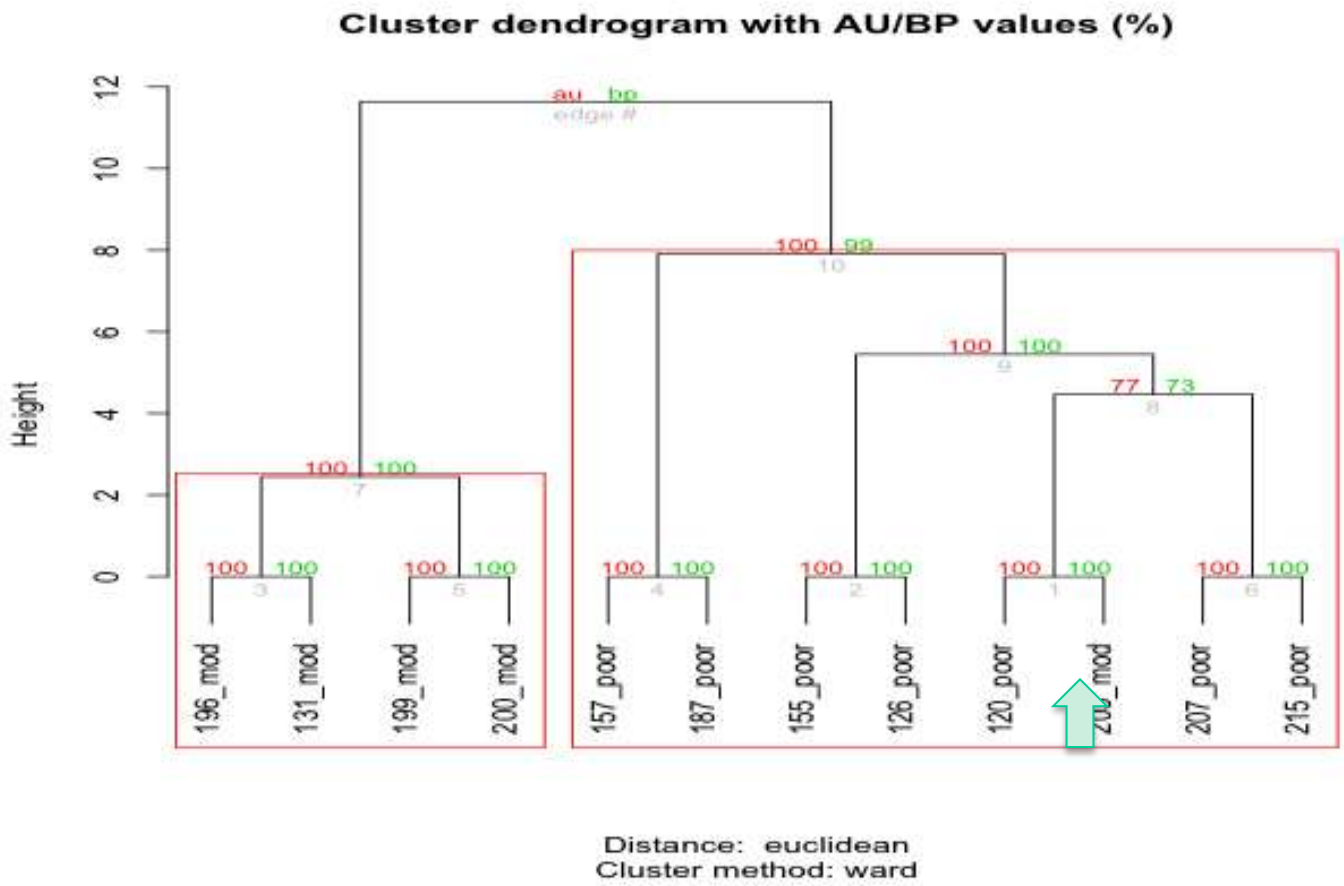
Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 1st revision.

Applying PSP to a HCC Dataset



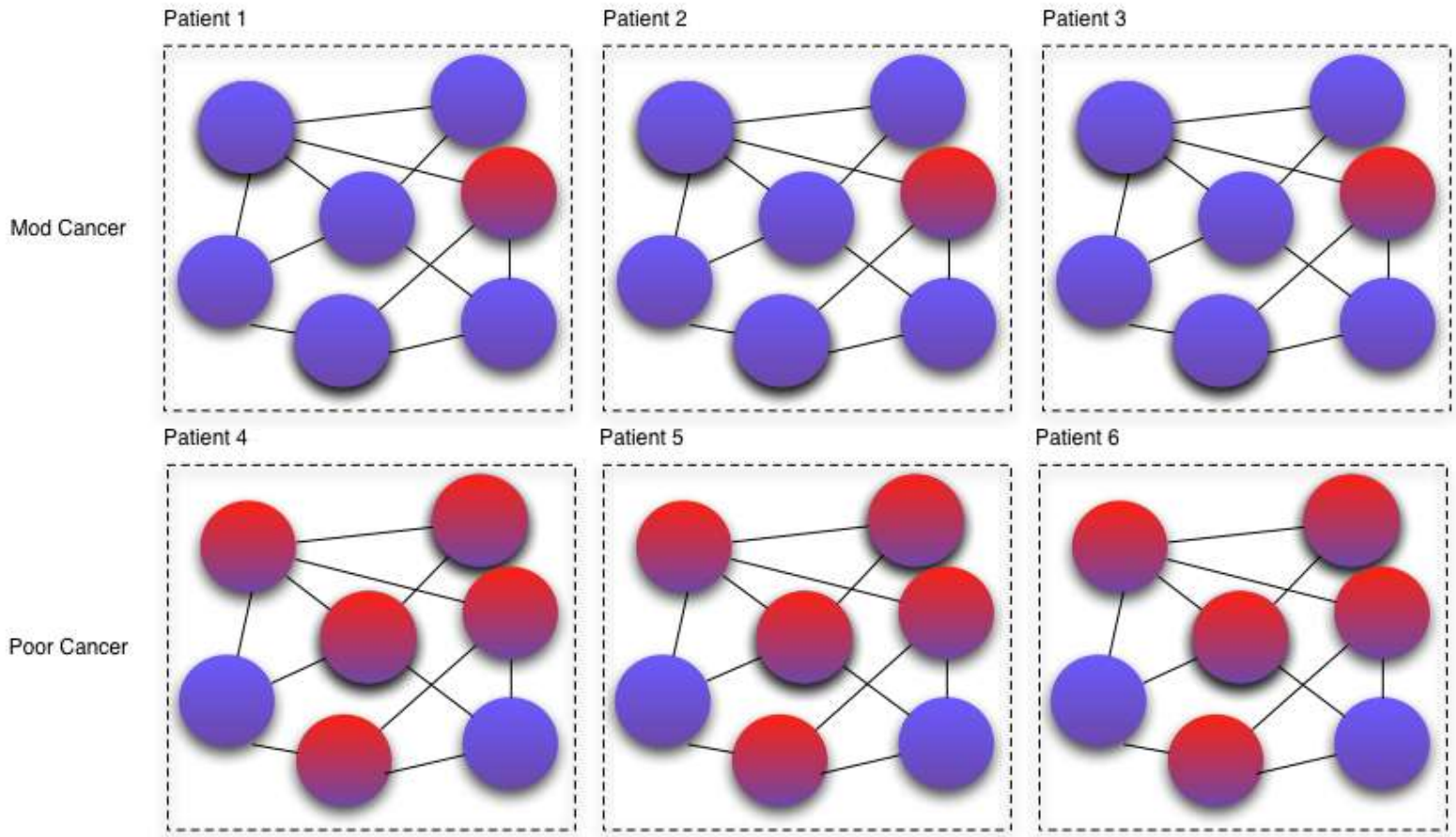
Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 1st revision.

Consistency: Samples segregate by their classes with high confidence



Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 1st revision.

Feature Selection



Detected Protein

Undetected Protein

$$t\_score = \frac{\bar{HA} - \bar{HB}}{S_{HA,HB}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where

$$S_{HA,HB} = \sqrt{\frac{(m-1)S_{HA}^2 + (n-1)S_{HB}^2}{m+n-2}}$$

Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 1st revision.

Top-Ranked Complexes

Cluster_ID	p_val	mod_score	poor_score	cluster_name
5179	0.000300541	0.513951977	3.159758312	NCOA6-DNA-PK-Ku-PARP1 complex
5235	0.000300541	0.513951977	3.159758312	WRN-Ku70-Ku80-PARP1 complex
1193	0.000300541	0.513951977	3.159758312	Rap1 complex
159	0	0	2.810927655	Condensin I-PARP-1-XRCC1 complex
2657	0.008815869	0	2.55616281	ESR1-CDK7-CCNH-MNAT1-MTA1-HDAC2 complex
3067	0.00911641	0	2.55616281	RNA polymerase II complex, incomplete (CDK8 complex), chromatin structure modifying
1226	0.013323983	0.715352108	2.420592827	H2AX complex I
5176	0	0.513951977	2.339059313	MGC1-DNA-PKcs-Ku complex
1189	0	0.513951977	2.339059313	DNA double-strand break end-joining complex
5251	0	0.513951977	2.339059313	Ku-ORC complex
2766	0	0.513951977	2.339059313	TERF2-RAP1 complex

Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 1st revision.

Top-Ranked GO Terms

GO ID	Description	No. of clusters
GO:0016032	viral reproduction	36
GO:0000398	nuclear mRNA splicing, via spliceosome	34
GO:0000278	mitotic cell cycle	28
GO:0000084	S phase of mitotic cell cycle	28
GO:0006366	transcription from RNA polymerase II promoter	26
GO:0006283	transcription-coupled nucleotide-excision repair	22
GO:0006369	termination of RNA polymerase II transcription	22
GO:0006284	base-excision repair	21
GO:0000086	G2/M transition of mitotic cell cycle	21
GO:0000079	regulation of cyclin-dependent protein kinase activity	20
GO:0010833	telomere maintenance via telomere lengthening	20
GO:0033044	regulation of chromosome organization	19
GO:0006200	ATP catabolic process	18
GO:0042475	odontogenesis of dentine-containing tooth	18
GO:0034138	toll-like receptor 3 signaling pathway	17
GO:0006915	apoptosis	17
GO:0006271	DNA strand elongation involved in DNA replication	17

A Shortcoming of PSP

- **Protein complex databases are still relatively small & incomplete…**

⇒ **Augment the set of protein complexes by protein clusters predicted from PPI networks!**

- **Many protein complex prediction methods**
  - CFinder, Adamcsek et al. *Bioinformatics*, 22:1021--1023, 2006
  - CMC, Liu et al. *Bioinformatics*, 25:1891--1897, 2009
  - CFA, Habibi et al. BMC Systems Biology, 4:129, 2010
  - …

Another Shortcoming of PSP

- **Protein complexes provided a biologically-rich feature set for PSP**
  - But it is only one aspect of biological function
- **The other aspect is biological pathways**
  - But coverage issue of proteomic profiles create lots of “holes”
- **Can we extract and use subnets from pathways?**

Another adaptation of SNet on proteomics profiles…

“Pathway-Derived Subnets” (PDS)

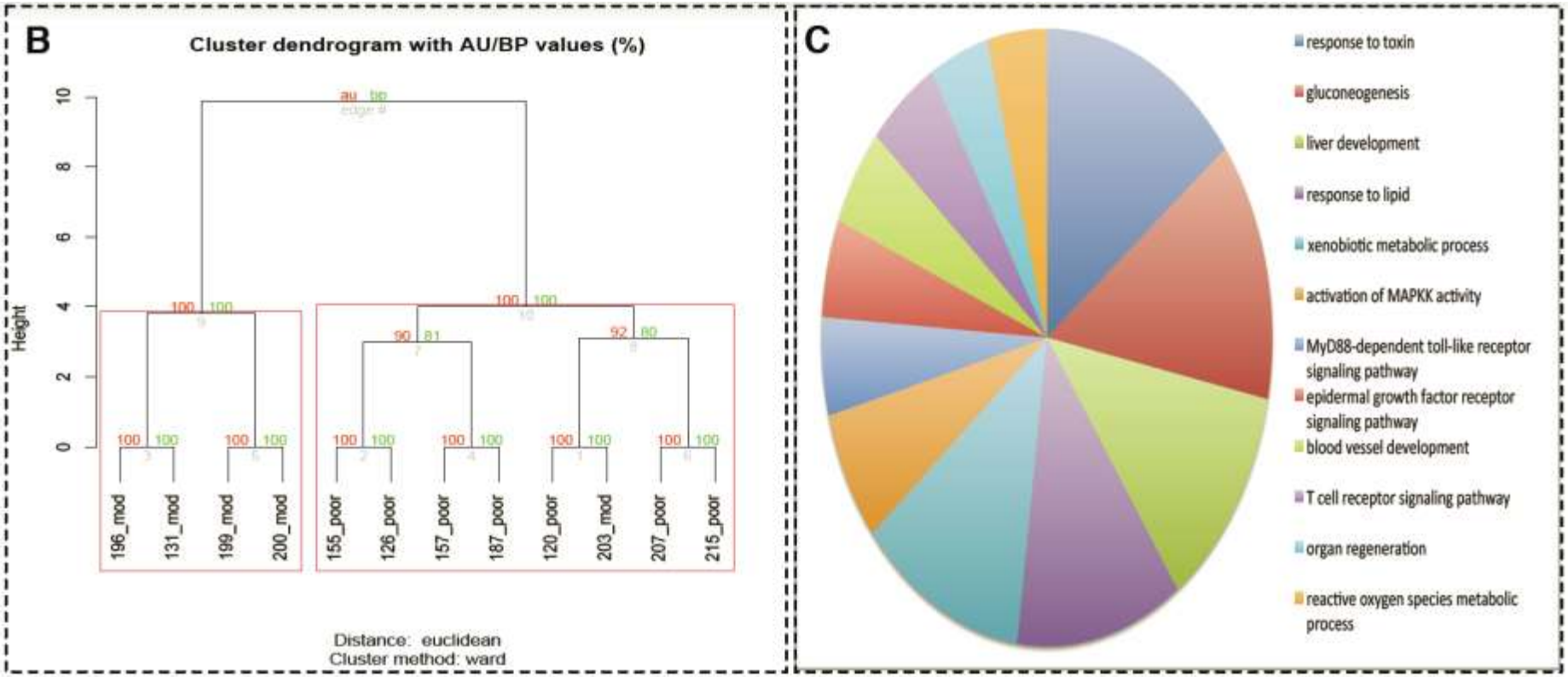
Pathway-Derived Subnets (PDS)

- **Identify the set S_i of proteins detected in more than 50% of samples having phenotype P_i**
  - Do this for each phenotype $P_1, \ldots, P_k$
- **Overlay $\cup_i S_i$ to pathways**
- **Remove nodes not covered by $\cup_i S_i$**

  $\Rightarrow$This fragments pathways into subnets
- **Use these subnets to form "proteomic signature profiles"**
  - The rest of the steps is same as PSP

Source: Wilson Goh

PDS consistently segregates mod vs poor patients



Source: Wilson Goh

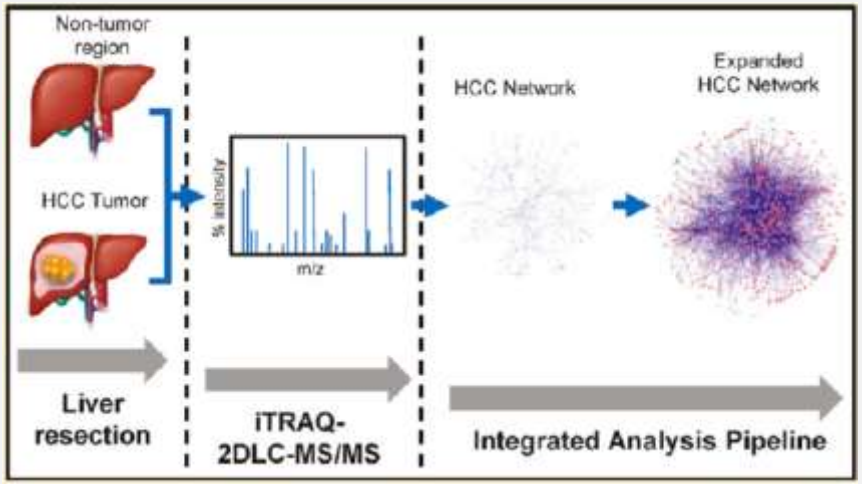
What have we learned?

- **Contextualization (into complexes and pathways) can deal with consistency issues in proteomics**

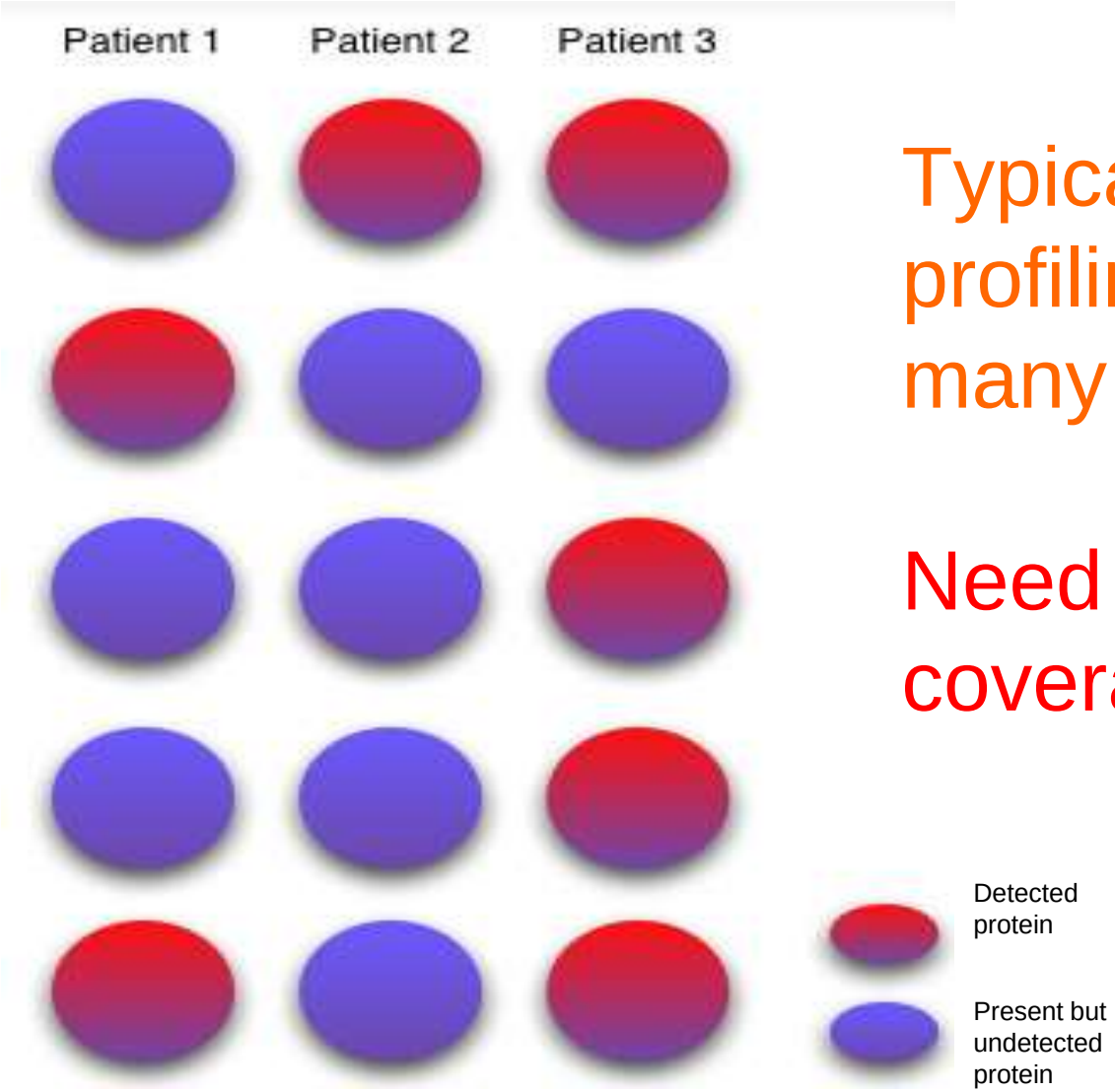
- **GO term analysis also indicates that context-based methods (PSP, PDS) select clusters that play integral roles in cancer**

- **Context-based methods (PSP, PDS) reveal many potential clusters and are not constrained by any prior arbitrary filtering which is a common first step in conventional analytical approaches**

Part 2: Delivering more powerful proteomic profile analysis



- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP
  - PDS
- **Improving coverage**
  - CEA
  - PEP
  - Max Link



Typical proteomic profiling misses many proteins

Need to improve coverage!

Image credit: Wilson Goh

Basic Approach

- **Rescue undetected proteins from high-scoring protein complexes**

- **Why?**

  Let A, B, C, D and E be the 5 proteins that function as a complex and thus are normally correlated in their expression. Suppose only A is not detected and all of B–E are detected. Suppose the screen has 50% reliability. Then, A's chance of being false negative is 50%, & the chance of B–E all being false positives is $(50\%)^4$=6%. Hence, it is almost 10x more likely that A is false negative than B–E all being false positives.

- **Shortcoming: Databases of known complexes are still small**

Li et al. **Network-assisted protein identification and data interpretation in shotgun proteomics**. *Mol. Syst. Biol., 5:*303, 2009.

CEA

- **Generate cliques from PPIN**
- **Rescue undetected proteins from cliques with containing many high-confidence proteins**

- **Reason: Cliques in a PPIN often correspond to proteins at the core of complexes**

- **Shortcoming: Cliques are too strict**

$\Rightarrow$ **Use more power complex prediction methods**

Goh et al. **A Network-based pipeline for analyzing MS data---An application towards liver cancer**. *Journal of Proteome Research*, 10(5):2261--2272, May 2011

PEP

- **Map high-confidence proteins to PPIN**
- **Extract immediate neighbourhood & predict protein complexes using CFinder**
- **Rescue undetected proteins from high-ranking predicted complexes**

- **Reason: Exploit powerful protein complex prediction methods**

- **Shortcoming: Hard to predict protein complexes**
  - Do we need to know all the proteins a complex?

Goh et al. **A Network-based maximum-link approach towards MS.** APBC 2012

MaxLink

- **Map high-confidence proteins ("seeds") to PPIN**
- **Identify proteins that talk to many seeds but few non-seeds**
- **Rescue these proteins**

- **Reason: Proteins interacting with many seeds are likely to be part of the same complex as these seeds**

- **Shortcoming: Likely to have more false-positives**

"Validation" of Rescued Proteins

- **Direct validation**
  - Use the original mass spectra to verify the quality of the corresponding y- and b-ion assignments
  - Immunological assay, etc.

- **Indirect validation**
  - Check whether recovered proteins have GO terms that are enriched in the list of seeds
  - Check whether recovered proteins show a pattern of differential expression betw disease vs normal samples that is similar to that shown by the seeds

An example using the PEP approach to recover undetected proteins …

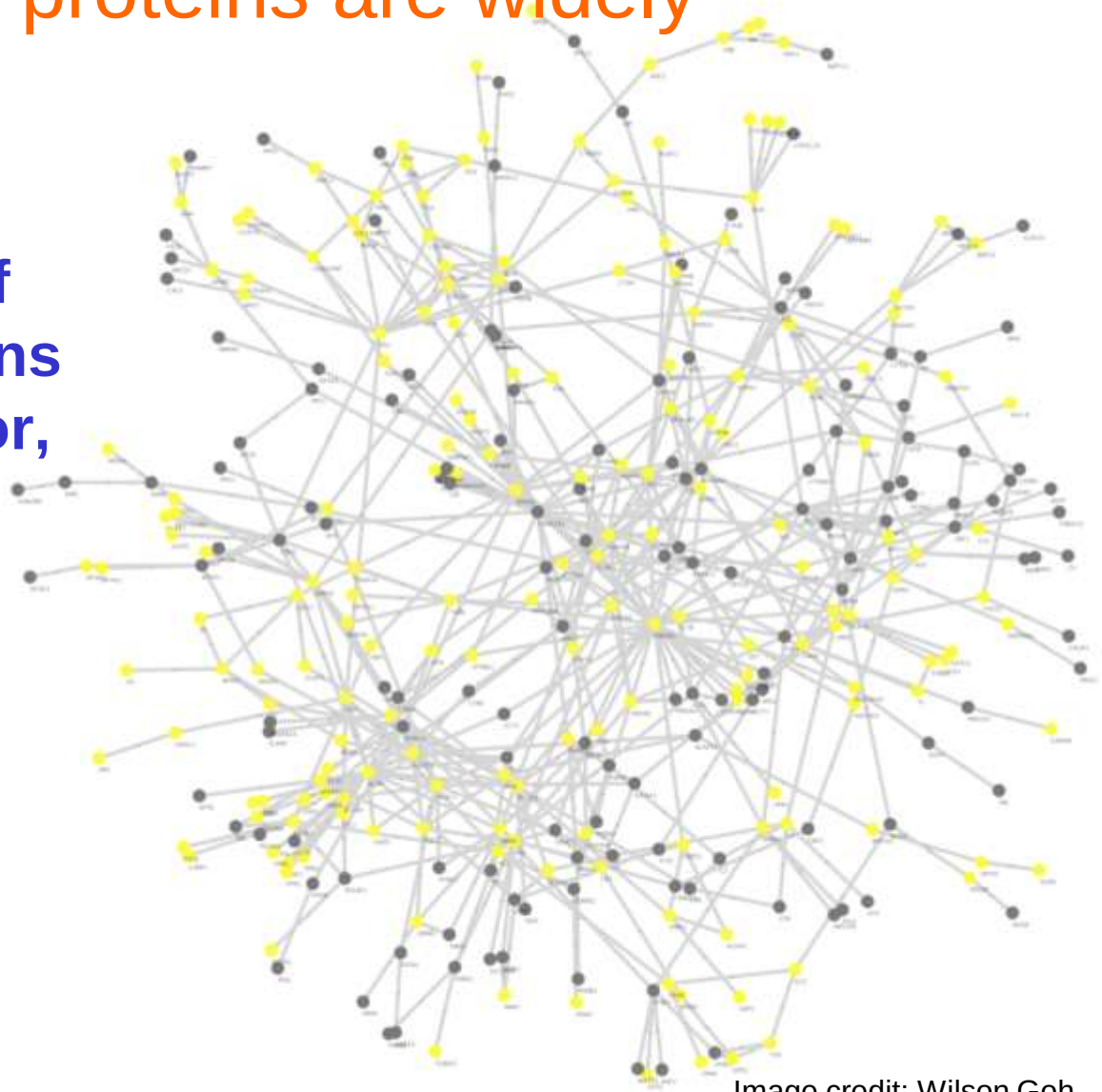
Background

- **HCC (Hepatocellular carcinoma)**
  - Classified into 3 phases: differentiated, moderately differentiated and poorly differentiated

- **Mass Spectrometry**
  - iTRAQ (Isobaric Tag for Relative and Absolute Quantitation)
  - Coupled with 2D LC MS/MS
  - Popular because of ability to run 8 concurrent samples in one go

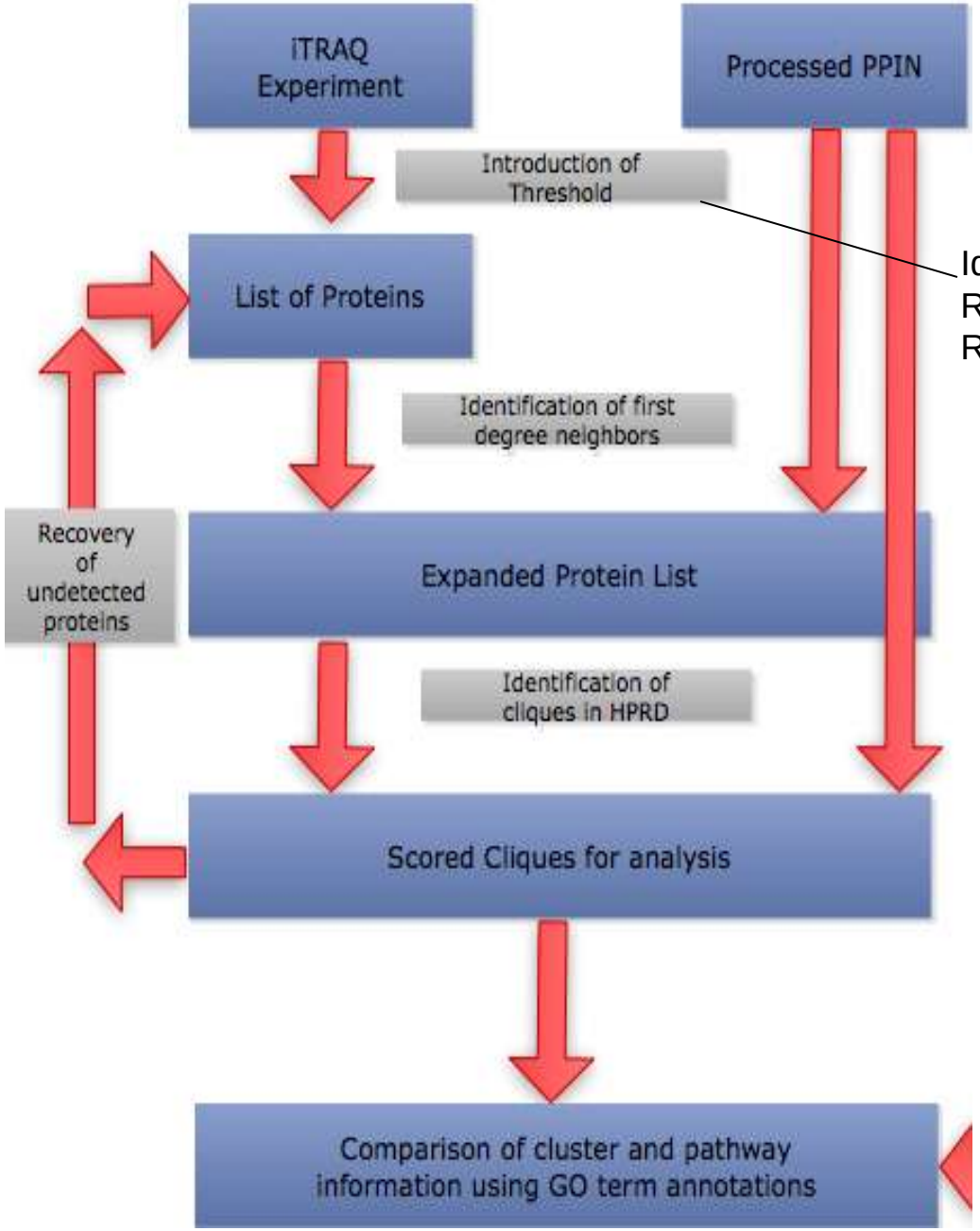
Poor and mod proteins are widely interspersed

- **In the subnet of reported proteins in mod and poor, poor and mod genes are well mixed**



- Mod and Poor
- Poor only

Image credit: Wilson Goh



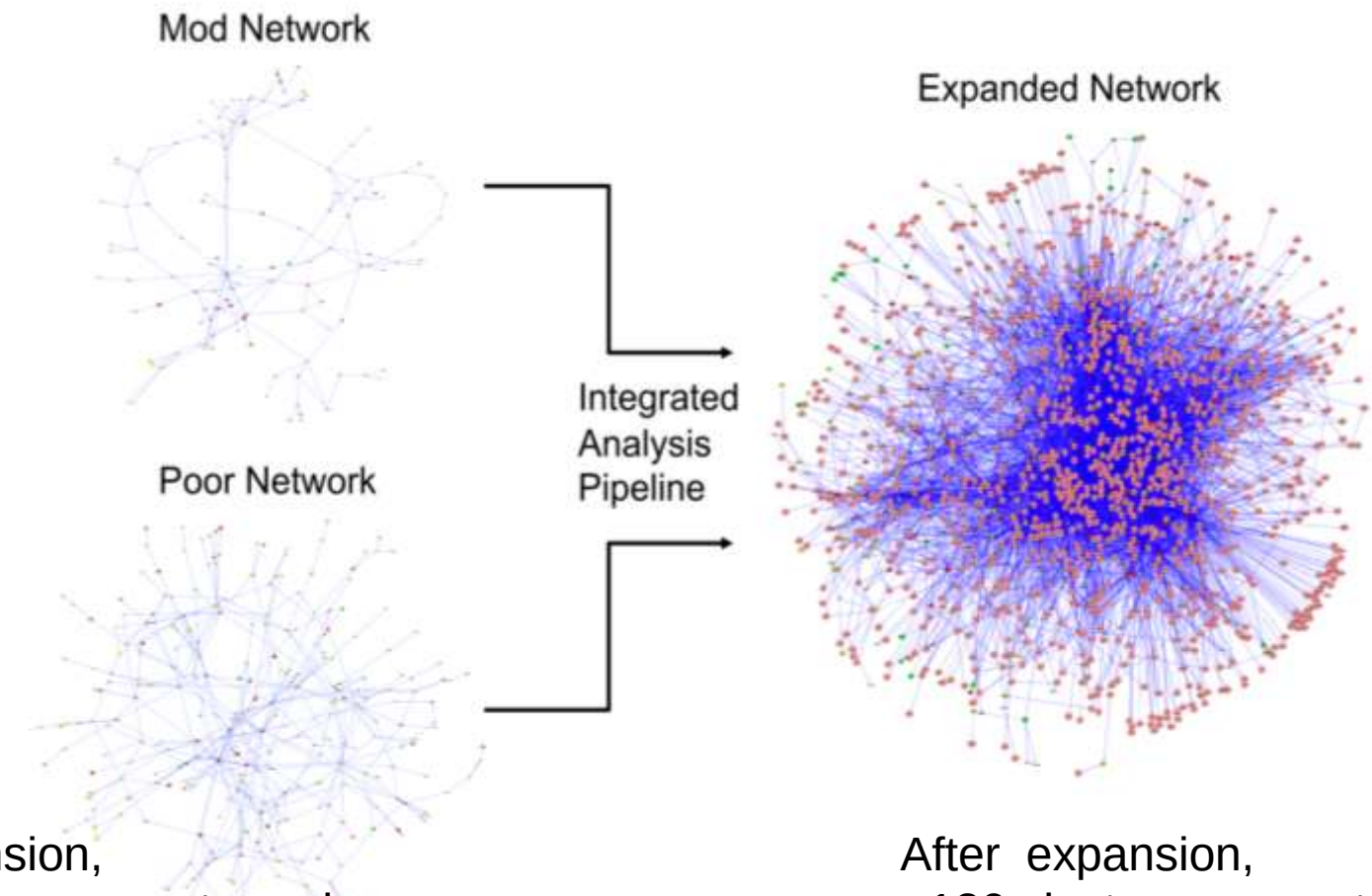
Identify the "seeds"
Ratio < 0.8 and > 1.25 for Mod (min 3 patients)
Ratio < 0.8 and > 1.25 for Poor (min 4 patients)

PEP Workflow

Goh et al. A Network-based pipeline for analyzing MS data---An application towards liver cancer. *Journal of Proteome Research*, 10(5):2261--2272, 2011

Goh et al. A Network-based pipeline for analyzing MS data---An application towards liver cancer. *Journal of Proteome Research*, 10(5):2261--2272, 2011

Expansion to include neighbors greatly improves coverage



W/o expansion,
4 k3 cliques were returned

After expansion,
~120 clusters were returned

Returning to Mass Spectra

- **Test set: Several proteins (ACTR2, CDC42, GNB2L1, KIF5B, PPP2R1A, PKACA and TOP1) from top 34 clusters not detected by Paragon**

- **The test: Examine their GPS and Mascot search results and their MS/MS-to-peptide assignments**

- **Assessment of MS/MS spectra of their top ranked peptides revealed accurate y- and b-ion assignments and were of good quality ($p < 0.05$)**

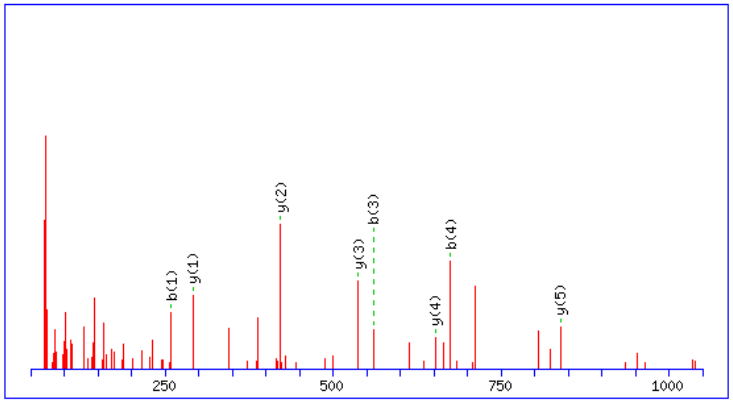
$\Rightarrow$ **In silico expansion verified**

Goh et al. *Journal of Proteome Research*, 10(5):2261--2272, 2011

Goh et al. *Journal of Proteome Research*, 10(5):2261--2272, 2011

Successful Verification

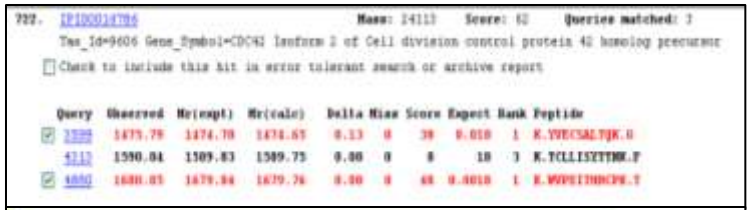
ACTR2

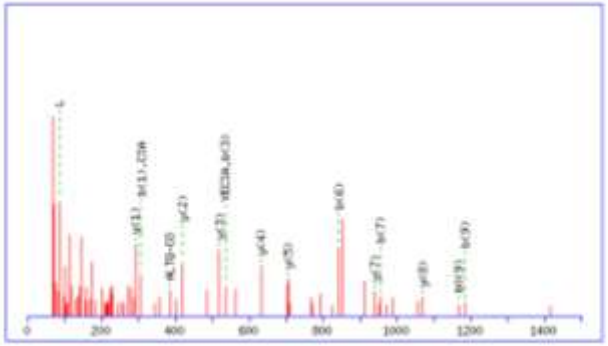


MONOISOTOPIC mass of neutral peptide Mr(calc): 1095.44
Fixed modifications: MMTS (C),(N-TERM)_iTRAQ,Lysine(K)_iTRAQ
Ions Score: 39 **Expect:** 0.018
Matches (Bold Red): 8/57 fragment ions using 15 most intense peaks

#	Immon.	a	a*	a^0	b	b*	b^0	Seq.	y	y*	y^0	#
1	87.06	231.16	214.13		**259.15**	242.13		N				6
2	159.09	417.24	400.21		445.23	428.21		W	**838.30**	821.27	820.29	5
3	88.04	532.26	515.24	514.25	**560.26**	543.23	542.25	D	**652.22**	635.19	634.21	4
4	88.04	647.29	630.26	629.28	**675.29**	658.26	657.28	D	**537.19**	520.17	519.18	3
5	104.05	778.33	761.30	760.32	806.33	789.30	788.32	M	**422.17**	405.14		2
6	245.12							K	**291.13**	274.10		1

CDC42



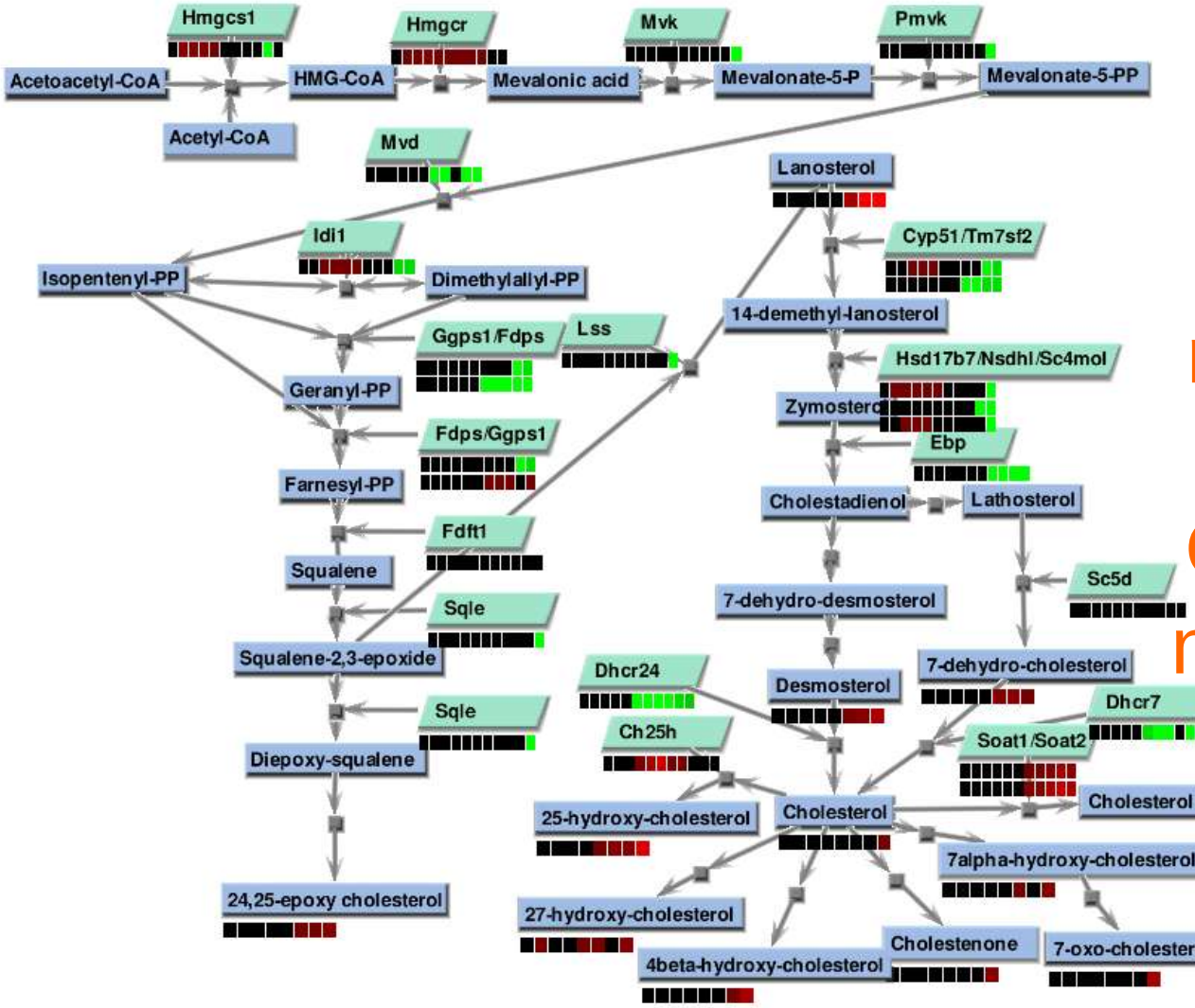


MONOISOTOPIC mass of neutral peptide Mr(calc): 1474.65
Fixed modifications: MMTS (C),(N-TERM)_iTRAQ,Lysine(K)_iTRAQ
Ions Score: 30 **Expect:** 0.018
Matches (Bold Red): 17/119 fragment ions using 26 most intense peaks

#	Immon.	a	a*	a^0	b	b*	b^0	Seq.	y	y*	y^0	#
1	136.08	280.18			**308.17**			Y				10
2	72.08	379.25			407.24			V	1168.49	1151.47	1150.48	9
3	102.05	508.29		490.28	**536.28**		518.27	E	**1069.42**	1052.40	1051.41	8
4	122.01	657.29		639.28	685.28		667.27	C	**940.38**	923.36	922.37	7
5	60.04	744.32		726.31	772.31		754.30	S	791.38	774.36	773.37	6
6	44.05	815.36		797.34	**843.35**		825.34	A	**704.35**	687.33	686.34	5
7	**86.10**	928.44		910.43	**956.43**		938.42	L	**633.32**	616.29	615.30	4
8	74.06	1029.49		1011.48	1057.48		1039.47	T	**520.23**	503.20	502.22	3
9	101.07	1157.55	1140.52	1139.53	**1185.54**	1168.51	**1167.53**	Q	**419.18**	402.16		2
10	245.12							K	**291.13**	274.10		1

References

- Käll & Vitek. **Computational Mass Spectrometry–Based Proteomics**. *PLoS Comput Biol* , 7(12): e1002277, 2011
- Goh et al. **How advancement in biological network analysis methods empowers proteomics**. *Proteomics*, in press
- [PSP] Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*, in press
- [CEA] Li et al. **Network-assisted protein identification and data interpretation in shotgun proteomics**. *Mol. Syst. Biol., 5:*303, 2009.
- [PEP] Goh et al. **A Network-based pipeline for analyzing MS data---An application towards liver cancer**. *J Proteome Research*, 10(5):2261-2272, 2011
- [MaxLink] Goh et al. **A Network-based maximum-link approach towards MS**. *APBC 2012*



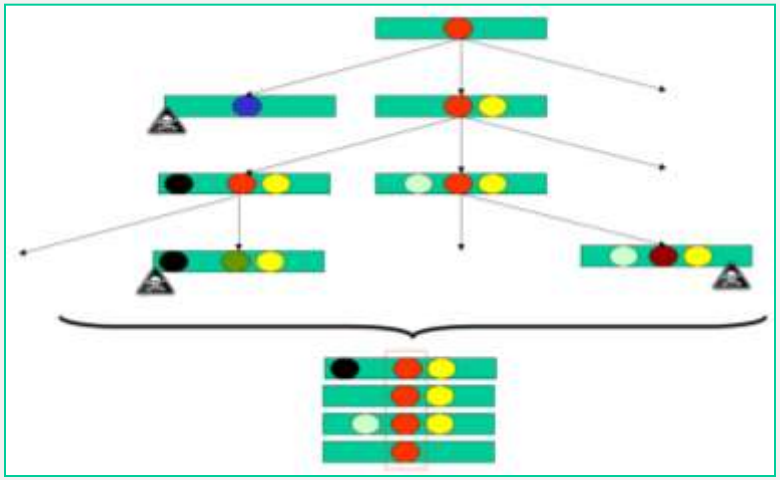
From proteomics to metabolomics & lipidomics: Can the same network-based approach be applied?

Using Biological Networks, Part 3: *Protein Function Prediction Without Informative Sequence Homologs*

Limsoon Wong

NUS
National University of Singapore

Part 3: Protein function prediction w/o informative sequence homologs



- **Basic protein function prediction**
- **"Guilt by association" of other properties**
- **Protein function prediction from PPIs**

A protein is a ...

- **A protein is a large complex molecule made up of one or more chains of amino acids**

- **Protein performs a wide variety of activities in the cell**

Function Assignment to Protein Sequence

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

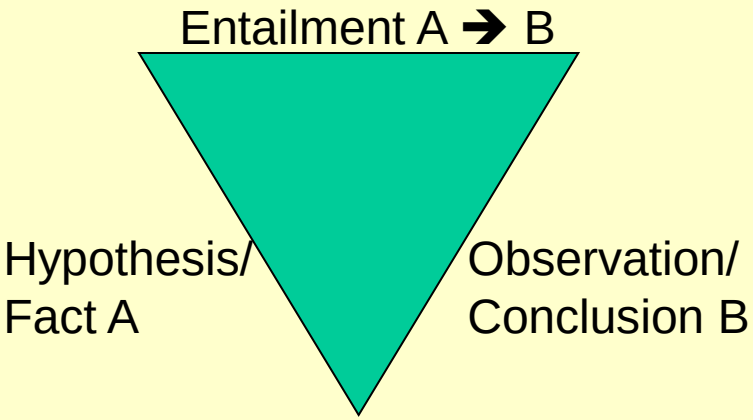
- **How do we attempt to assign a function to a new protein sequence?**

Invariant and Abductive Reasoning

- **Function is determined by 3D struct of protein & environment protein is in**
- **Constraints imposed by 3D struct & environment give rise to "invariant" properties observed in proteins having the ancestor with that function**

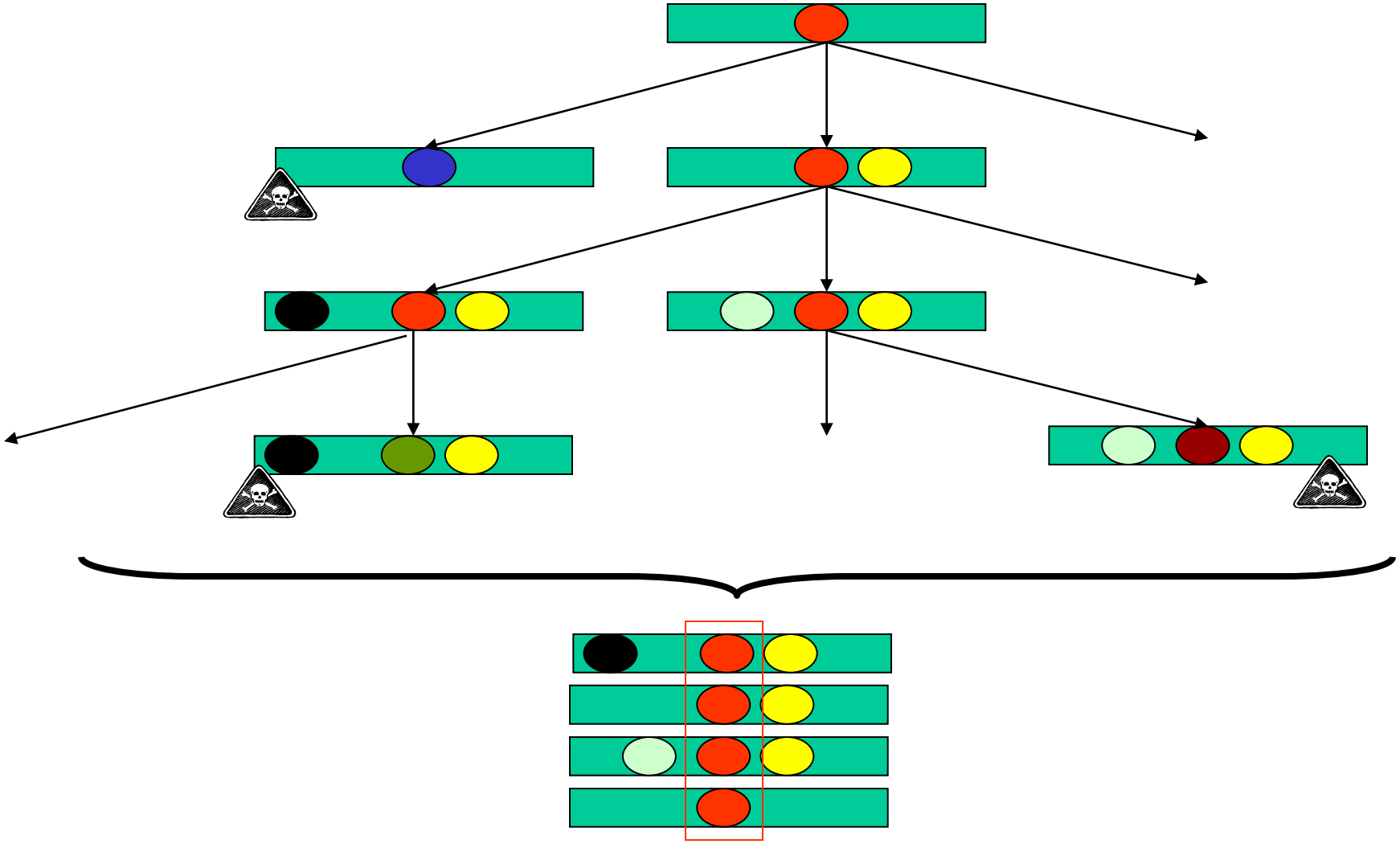
⇒ **Abductive reasoning**

– If those invariant properties are seen in a protein, then the protein is homolog of this protein



⇒ **"Guilt by association"**

In the course of evolution…



Guilt-by-Association

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

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

```
                          60        70        80        90       100
Amicyanin            MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTPTEAGTYDYHCTPHPFMRGKVVV
                                                   :.:  . ::. ::
Ascorbate Oxidase    ILQRGTPWADGTASISQCAINPGETFFYNPTVDNPGTFFYHGHLGMQRSAGLYG
                          70        80        90       100       110
```

No obvious match between Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

```
>gi|13476732|ref|NP_108301.1|   unknown protein [Mesorhizobium loti]
 gi|14027493|dbj|BAB53762.1|    unknown protein [Mesorhizobium loti]
          Length = 105

 Score =  105 bits (262), Expect = 1e-22
 Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIKWVNKDVFAHT 60
           MK G L  ++        MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSVLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between Amicyanin and unknown M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Guilt-by-Association: Caveats

- **Ensure that the effect of database size has been accounted for**
- **Ensure that the function of the homology is not derived via invalid "transitive assignment"**
- **Ensure that the target sequence has all the key features associated with the function, e.g., active site and/or domain**

Law of Large Numbers

- Suppose you are in a room with 365 other people
- Q: What is the prob that a specific person in the room has the same birthday as you?
- A: 1/365 = 0.3%
- Q: What is the prob that there is a person in the room having the same birthday as you?
- A: $1 - (364/365)^{365} = 63\%$
- Q: What is the prob that there are two persons in the room having the same birthday?
- A: 100%

Interpretation of P-value

- **Seq. comparison progs, e.g. BLAST, often associate a P-value to each hit**
- **P-value is interpreted as prob that a random seq has an equally good alignment**

- **Suppose the P-value of an alignment is 10^{-6}**
- **If database has 10^7 seqs, then you expect $10^7 * 10^{-6}$ = 10 seqs in it that give an equally good alignment**

$\Rightarrow$ **Need to correct for database size if your seq comparison prog does not do that!**

Note: $P = 1 - e^{-E}$

Exercise: Name a commonly used method for correcting p-value for a situation like this

Lightning Does Strike Twice!

- **Roy Sullivan, a former park ranger from Virgina, was struck by lightning 7 times**
  - 1942 (lost big-toe nail)
  - 1969 (lost eyebrows)
  - 1970 (left shoulder seared)
  - 1972 (hair set on fire)
  - 1973 (hair set on fire & legs seared)
  - 1976 (ankle injured)
  - 1977 (chest & stomach burned)

- **September 1983, he committed suicide**

Cartoon: Ron Hipschman
Data: David Hand

Effect of Seq Compositional Bias

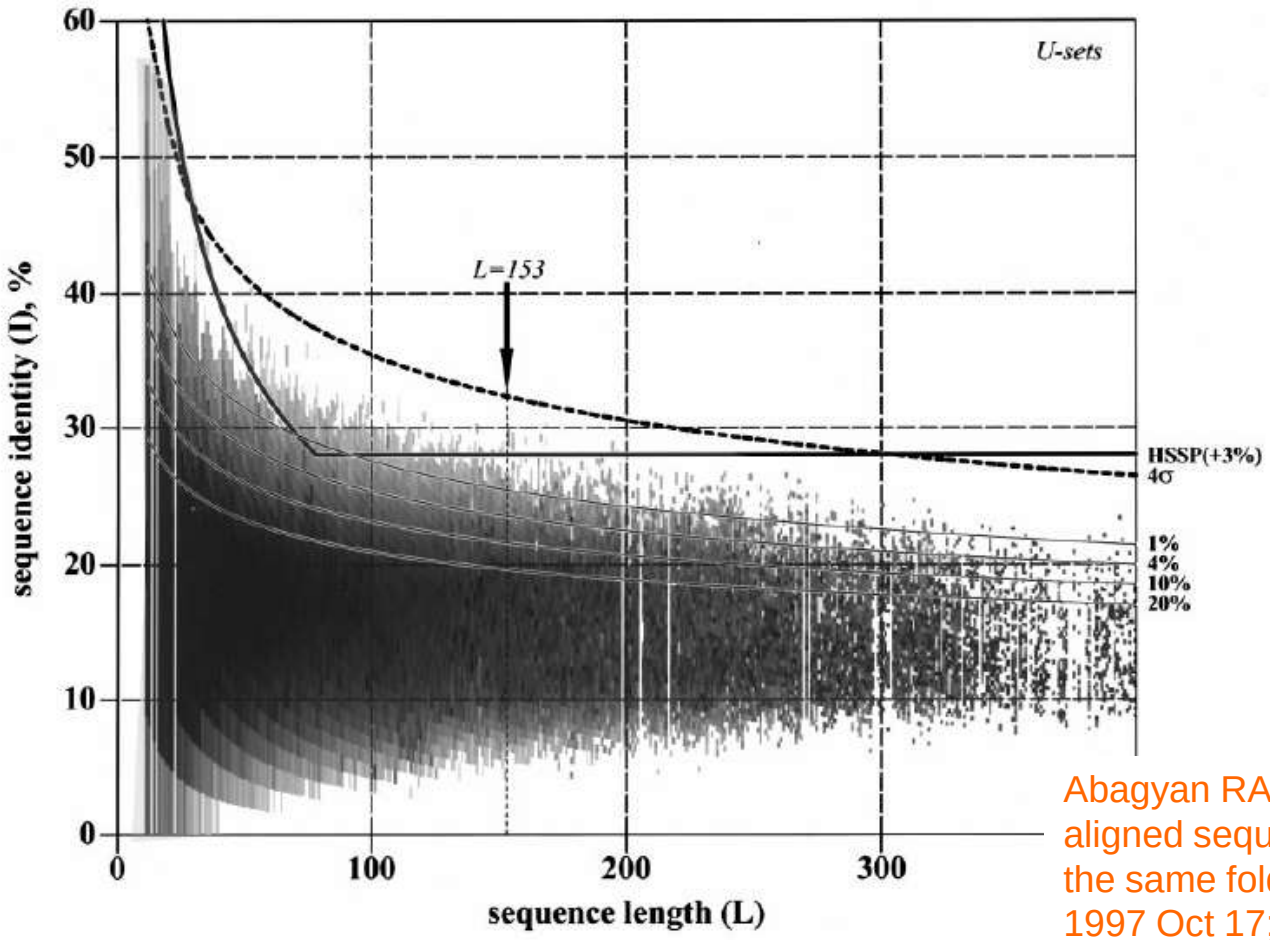
- **One fourth of all residues in protein seqs occur in regions with biased amino acid composition**
- **Alignments of two such regions achieves high score purely due to segment composition**

⇒ **While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments**

- **E.g., by default, BLAST employs the SEG algo to filter low complexity regions from proteins before executing a search**

Source: NCBI

Effect of Sequence Length



Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol. 1997 Oct 17;273(1):355-68

Examples of Invalid Function Assignment:

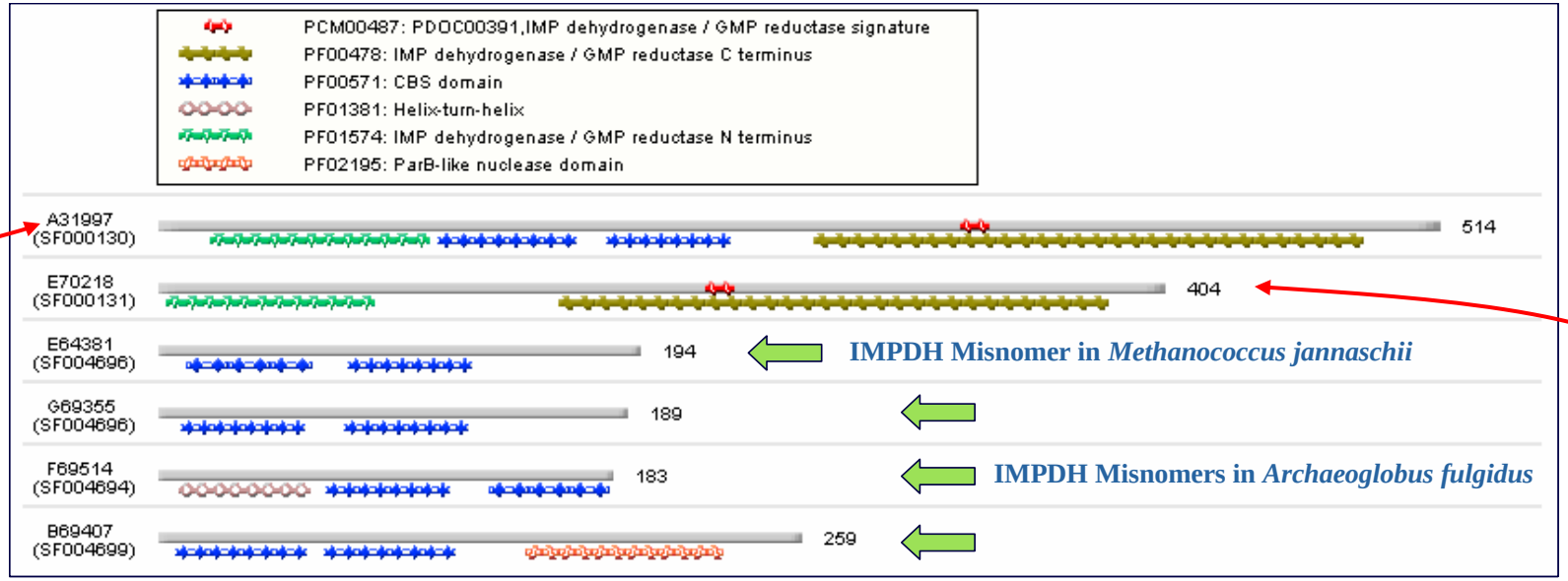
The IMP Dehydrogenases (IMPDH)

18 entries were found

ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
NF00181857	Methanococcus jannaschii	E64381 conserved hypothetical protein MJ0653	Y653_METJA Hypothetical protein MJ0653	g1592300 inosine-5'-monophosphate dehydrogenase (guaB) NP_247637 inosine-5'-monophosphate dehydrogenase (guaB)
NF00187788	Archaeoglobus fulgidus	G69355 MJ0653 homolog AF0847 *ALT_NAMES*: inosine-monophosphate dehydrogenase (guaB-1) homolog [misnomer]	O29411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)	g2649754 inosine monophosphate dehydrogenase (guaB-1) NP_069681 inosine monophosphate dehydrogenase (guaB-1)
NF00188267	Archaeoglobus fulgidus	F69514 yhcV homolog 2 *ALT_NAMES*: inosine-monophosphate dehydrogenase (guaB-2) homolog [misnomer]	O28162 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-2)	g2648410 inosine monophosphate dehydrogenase (guaB-2) NP_070943 inosine monophosphate dehydrogenase (guaB-2)
NF00188697	Archae			ophosphate ive nophosphate ive
NF00197776	Thermo			onophosphate d protein nonophosphate d protein
NF00414709	Methanothermobacter thermautotrophicus	[illegible] *ALT_NAMES*: inosine-monophosphate dehydrogenase related protein V [misnomer]	O27294 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	onophosphate dehydrogenase related protein V NP_276354 inosine-5'-monophosphate dehydrogenase related protein V
NF00414811	Methanothermobacter thermautotrophicus	D69035 MJ1232 protein homolog MTH126 *ALT_NAMES*: inosine-5'-monophosphate dehydrogenase related protein VII [misnomer]	O26229 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	g2621166 inosine-5'-monophosphate dehydrogenase related protein VII NP_275269 inosine-5'-monophosphate dehydrogenase related protein VII
NF00414837	Methanothermobacter thermautotrophicus	H69232 MJ1225-related protein MTH992 *ALT_NAMES*: inosine-5'-monophosphate dehydrogenase related protein IX [misnomer]	O27073 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	g2622093 inosine-5'-monophosphate dehydrogenase related protein IX NP_276127 inosine-5'-monophosphate dehydrogenase related protein IX
NF00414969	Methanothermobacter thermautotrophicus	B69077 yhcV homolog 2 *ALT_NAMES*: inosine-monophosphate dehydrogenase related protein X [misnomer]	O27616 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	g2622697 inosine-5'-monophosphate dehydrogenase related protein X NP_276687 inosine-5'-monophosphate dehydrogenase related protein X

A partial list of IMPdehydrogenase misnomers in complete genomes remaining in some public databases

IMPDH Domain Structure

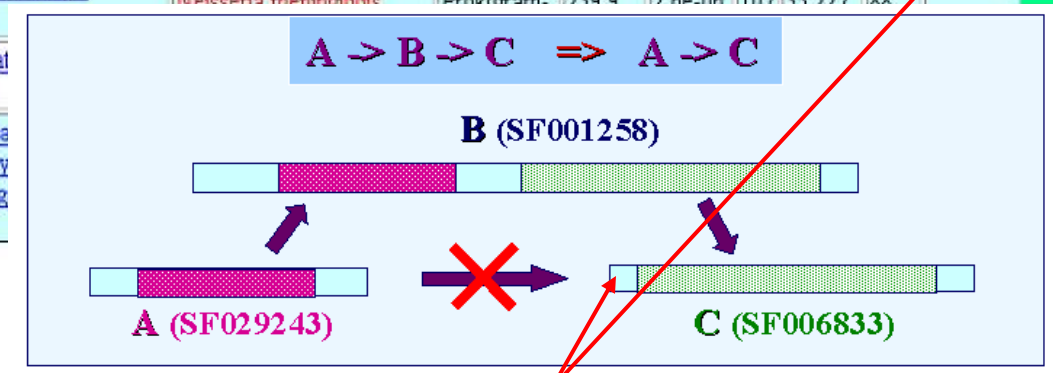


- **Typical IMPDHs have 2 IMPDH domains that form the catalytic core and 2 CBS domains.**
- **A less common but functional IMPDH (E70218) lacks the CBS domains.**
- **Misnomers show similarity to the CBS domains**

Invalid Transitive Assignment

Root of invalid transitive assignment

B	H70468	SF001258	051440	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Aquifex aeolicus	Prok/other	594.3	4.8e-26	205	39.086	197	
	S76963	SF001258	039935	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Synechocystis sp.	Prok/gram-	557.0	5.7e-24	230	39.175	194	
	T35073	SF029243	005738	probable phosphoribosyl-AMP cyclohydrolase	Streptomyces coelicolor	Prok/gram+	399.3	3.5e-15	128	42.157	102	
	S53349	SF001257	001188	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)	Saccharomyces cerevisiae	Euk/fungi	384.1	2.5e-14	799	31.863	204	
A	E69493	SF029243	005738	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) [similarity]	Archaeoglobus fulgidus	Archae	396.8	4.8e-15	108	47.778	90	
C	G64337	SF006833	030827	phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Methanococcus jannaschii	Archae	246.9	1.1e-06	95	36.842	95	
	D81178	SF006833	101491	phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMB0603 [similarity]	Neisseria meningitidis	Prok/gram-	239.9	2.6e-06	107	35.227	88	
	G81925	SF006833	101491	phosphoribosyl-ATP pyrophosphat 3.6.1.31) NMA0807 [similarity]								
	S51513	SF001257	001188	phosphoribosyl-AMP cyclohydrola 3.5.4.19) / phosphoribosyl-ATP py (EC 3.6.1.31) / histidinol dehydrog 1.1.1.23)								



Mis-assignment of function

No IMPDH domain

Part 3: Protein function prediction w/o informative sequence homologs

- Basic protein function prediction
- **"Guilt by association" of other properties**
- Protein function prediction from PPIs

What if there is no useful seq homolog?

- **Guilt by other types of association!**
  - Domain modeling (e.g., HMMPFAM)
  - ✓ Similarity of phylogenetic profiles
  - ✓ Similarity of dissimilarities (e.g., SVM-PAIRWISE)
  - 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 Phylogenetic Profiles

- **Proteins carry out their function within the context of biological pathways**

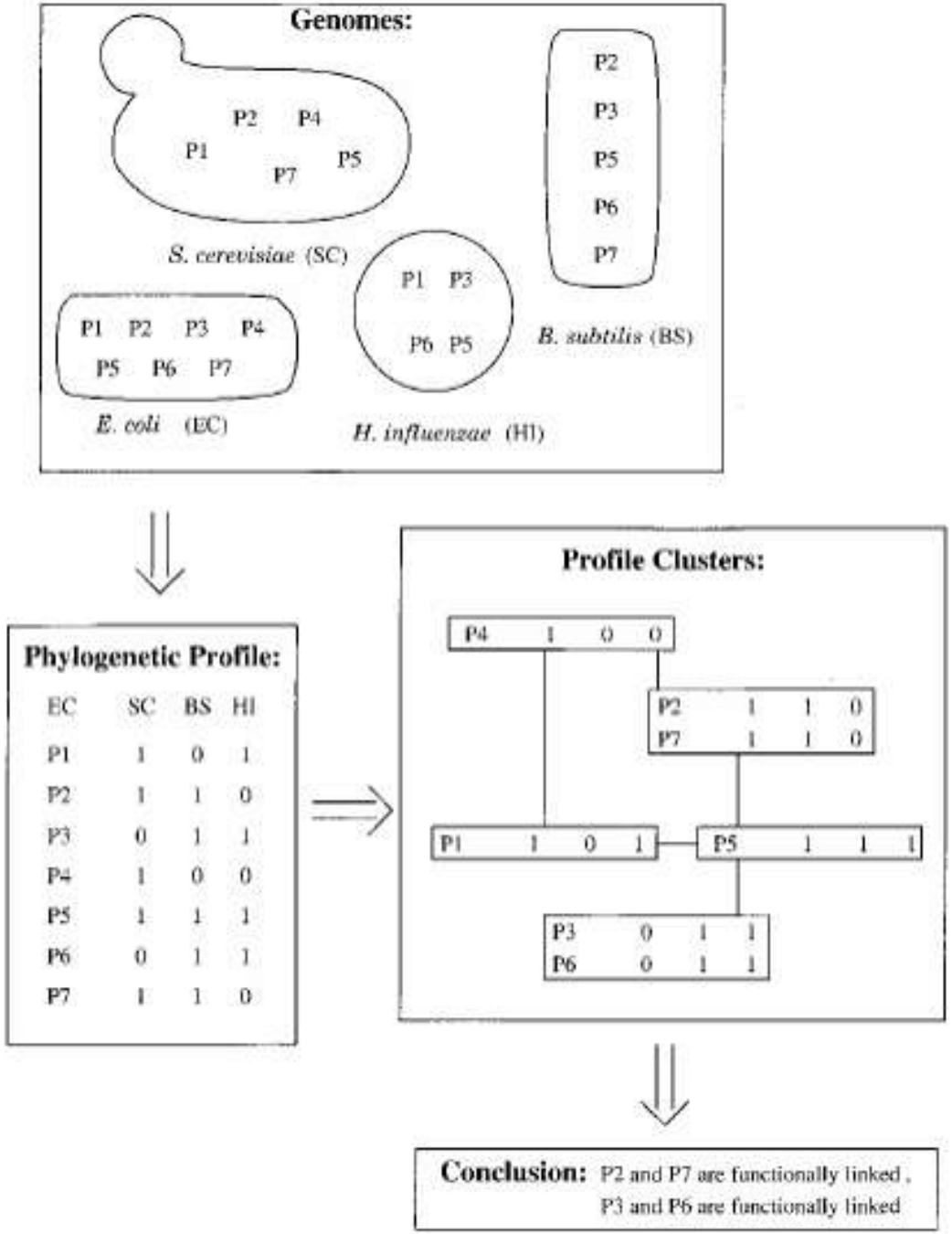
- **Genes coding for proteins participating in the same pathway are present together**

By abduction,

- **Genes (and hence proteins) with identical patterns of occurrence across phyla participate in the same pathway and function together**

$\Rightarrow$ **Phylogenetic profiling**

Phylogenetic Profiling: How it Works



Pellegrini et al., *PNAS*, 96:4285--4288, 1999

Phylogenetic Profiles: Evidence

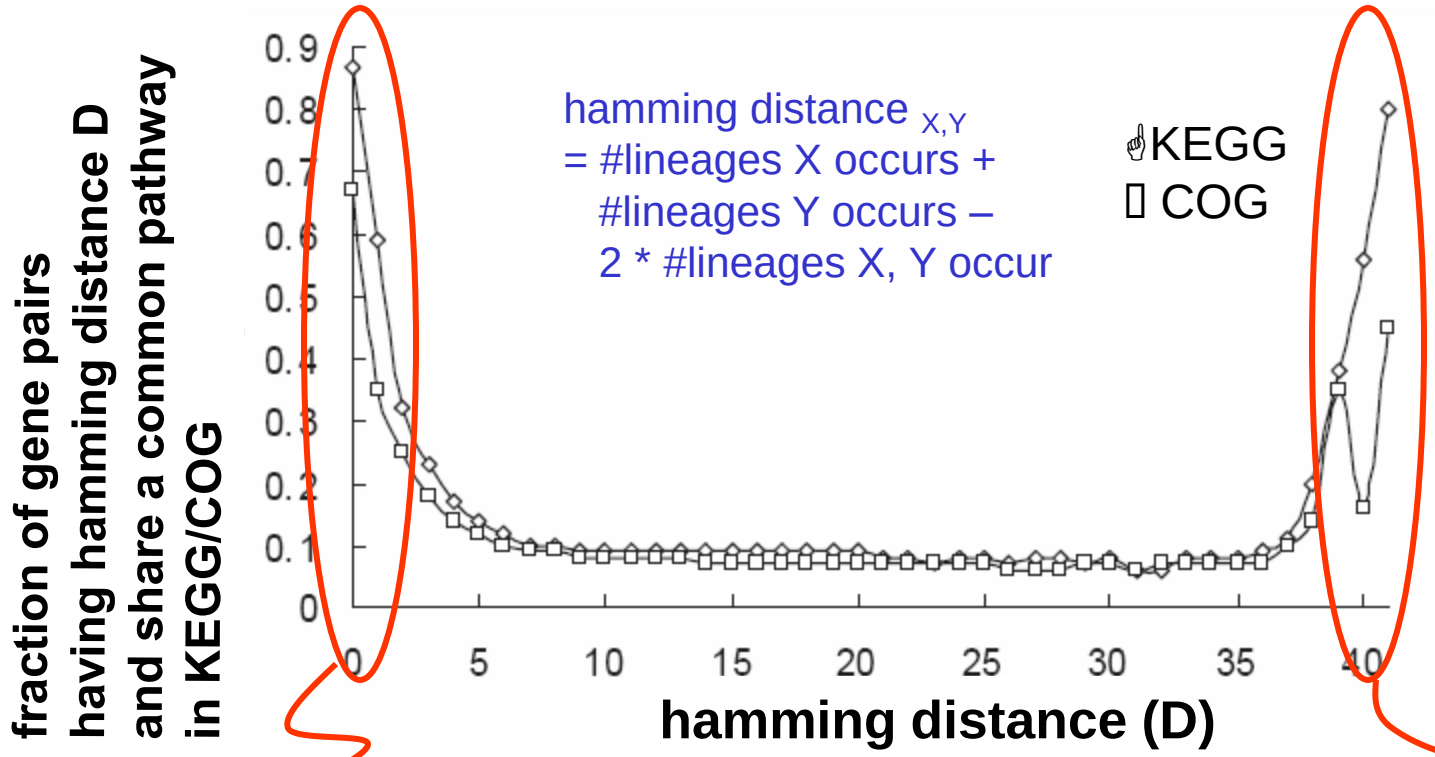
Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin and Molybdenum, and molybdoterin	12	6	1
Hypothetical†	1,084	108,226	8,440

- **E. coli proteins grouped based on similar keywords in SWISS-PROT have similar phylogenetic profiles**

Pellegrini et al., *PNAS*, 96:4285--4288, 1999

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

Phylogenetic Profiling: Evidence



- **Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways**

Why do proteins having high hamming distance also have this behaviour?

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**	...
...	...	...	...

Li & Noble. *JCB*, 10(6):857-868, 2003

SVM-Pairwise Framework

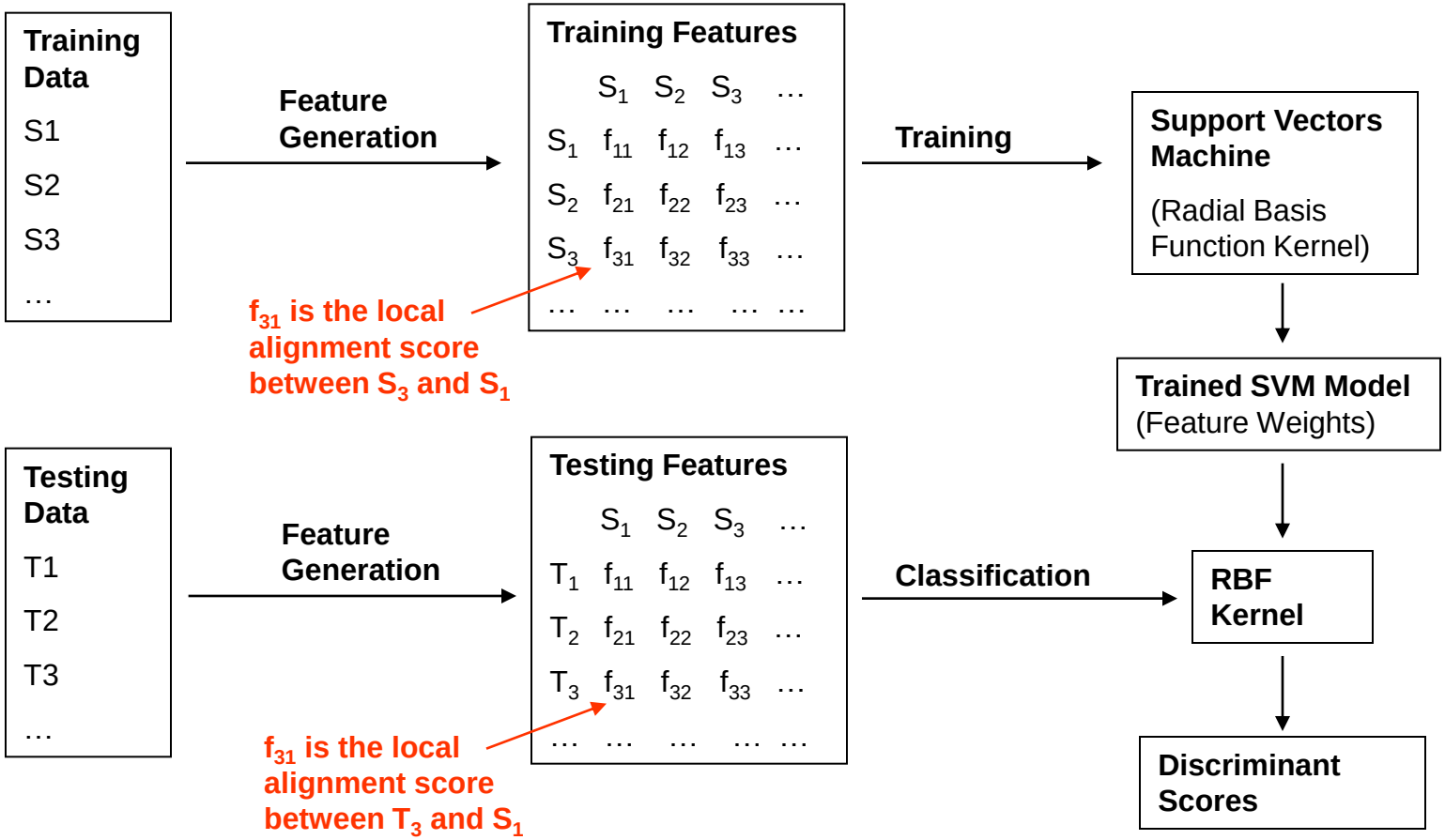
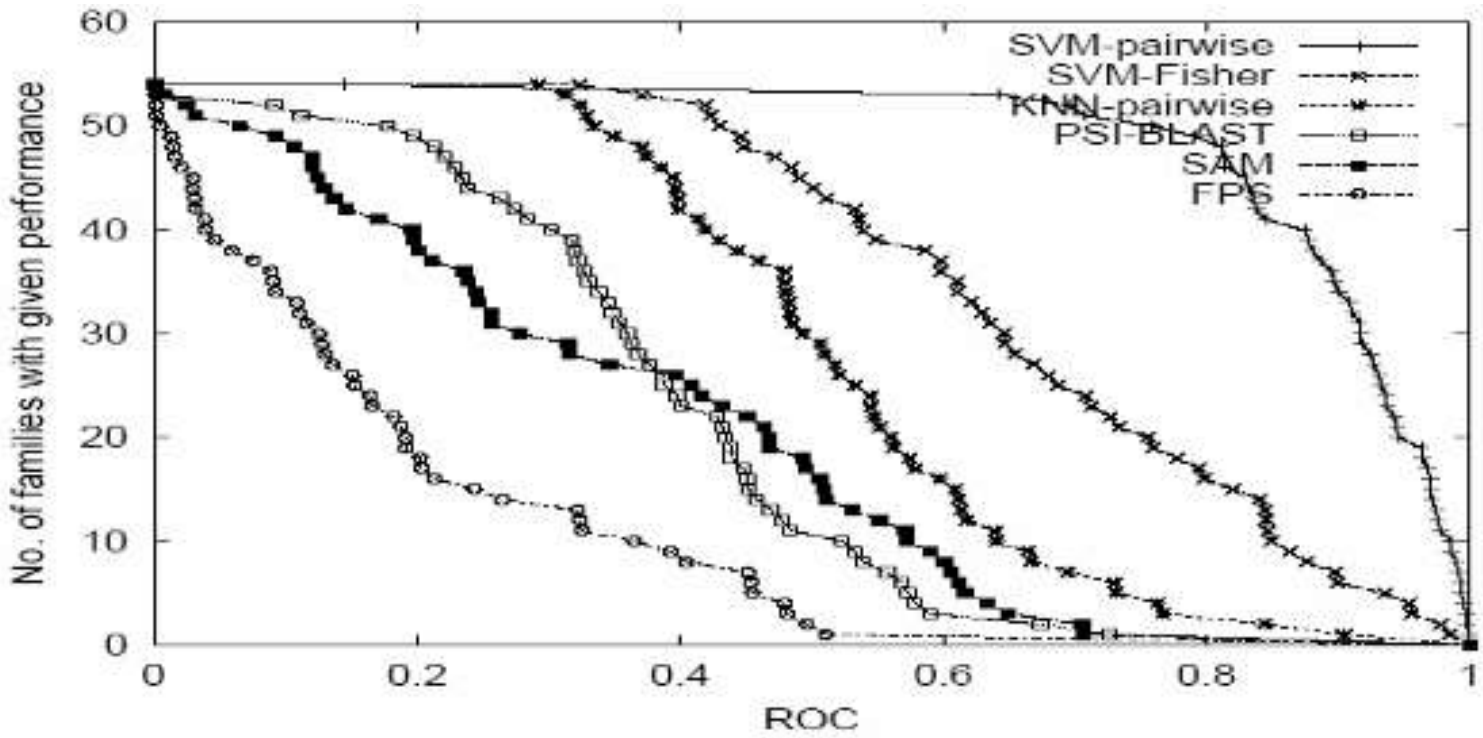


Image credit: Kenny Chua

Li & Noble. *JCB*, 10(6):857-868, 2003

Performance of SVM-Pairwise



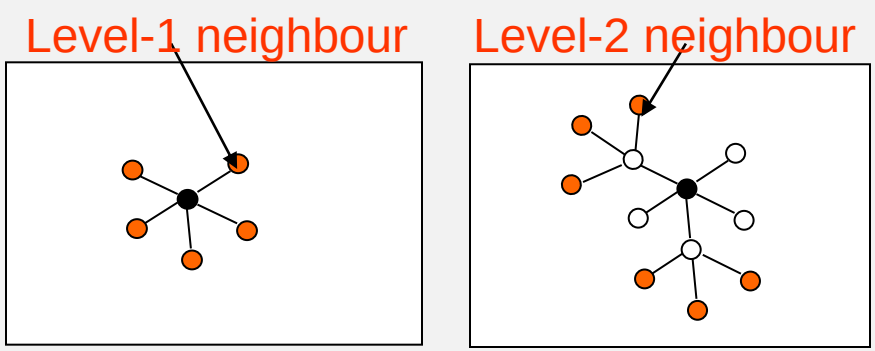
- **Receiver Operating Characteristic (ROC)**
  - The area under the curve derived from plotting true positives as a function of false positives for various thresholds.

References

- Hawkins & Kihara. **Function prediction of uncharacterized proteins**. *JBCB*, 5(1):1-30, 2007
- [Phylogenetic Profile] Pellegrini et al. **Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles**, *PNAS*, 96:4285-4288, 1999
- [Phylogenetic Profile] Wu et al. **Identification of functional links between genes using phylogenetic profiles**, *Bioinformatics*, 19:1524-1530, 2003
- [SVM-Fisher] Jaakkola et al. **A discriminative framework for detecting remote homologies**. *JCB*, 7(1-2):95-11, 2000
- [SVM-Pairwise] Li & Noble. **Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships.** *JCB*, 10(6):857-868, 2003

Part 3: Protein function prediction w/o informative sequence homologs

- Basic protein function prediction
- "Guilt by association" of other properties
- **Protein function prediction from PPIs**



Main Hypotheses of PPIN-Based Function Prediction

- **Proteins with similar function are topologically close in PPIN**
  - Direct functional association
  - Indirect functional association

A pair of proteins that participate in the same cellular processes or localize to the same cellular compartment are many times more likely to interact than a random pair of proteins

- **Proteins with similar function have interaction neighborhoods that are similar**

When proteins in the neighborhood of a protein X have similar functions to proteins in the neighborhood of a protein Y, then proteins X & Y likely operate in similar environment

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

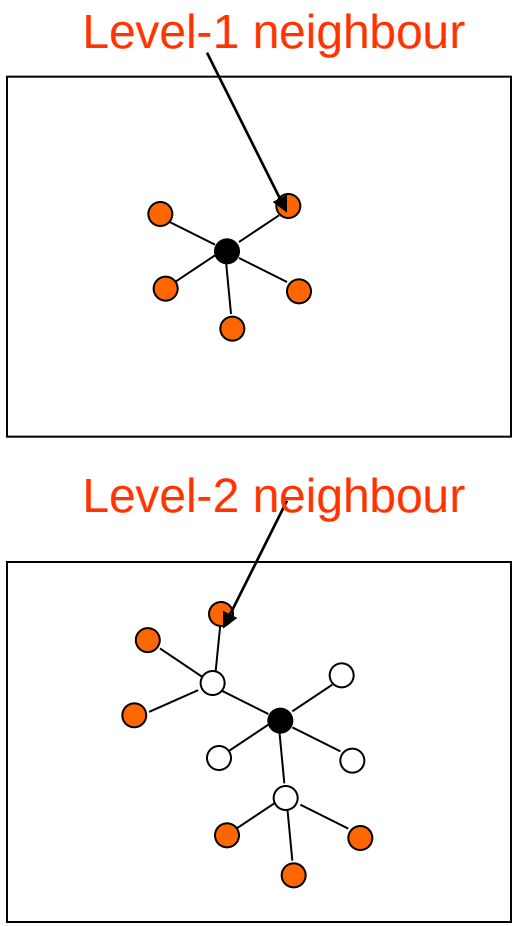
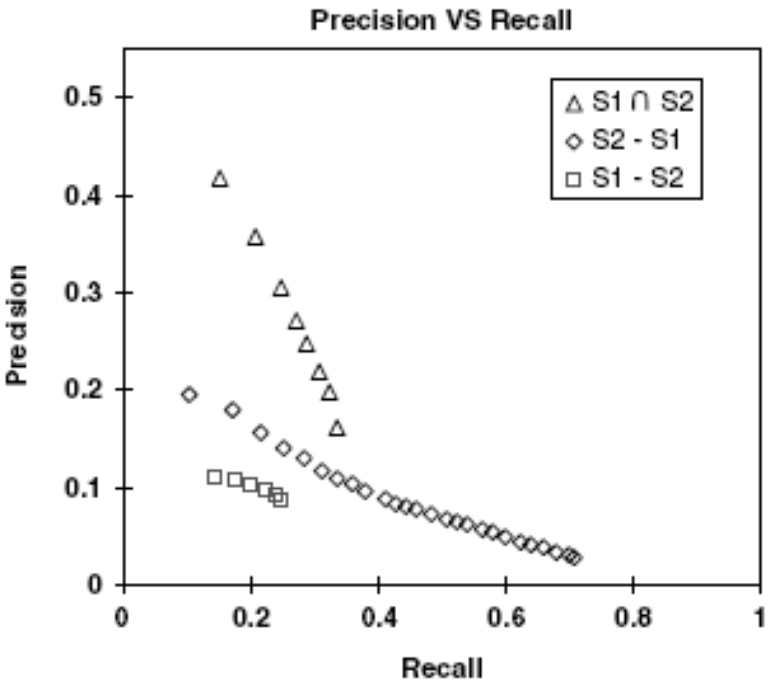


Image credit: Kenny Chua

Majority Voting

- **Proteins with similar function are topologically close in PPIN**

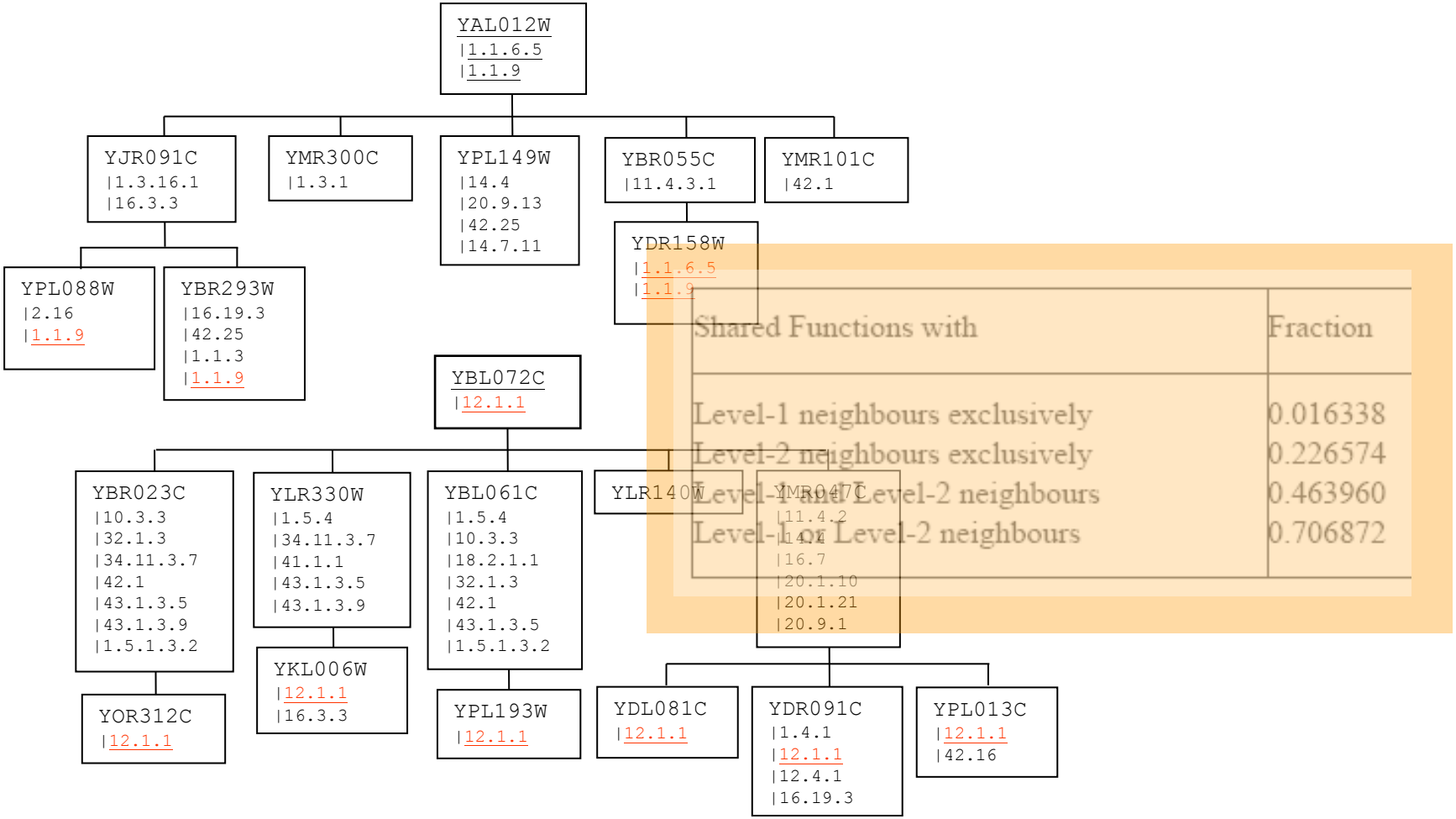
- **Assign a protein a function that is over represented among its interaction partners**



- **Shortcomings**
  - L1 is not sensitive
  - L2 is noisy

Hishigaki et al. *Yeast*, 18:523-531, 2001

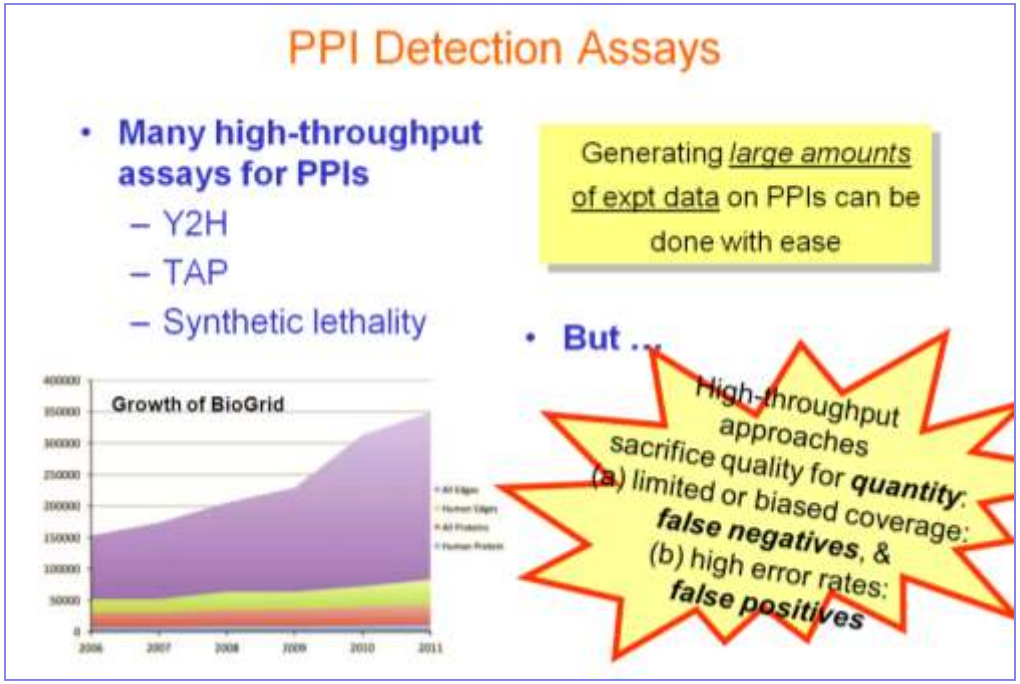
Why is L1 not sensitive?



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*, 22:1623-1630, 2006.

Why is L2 noisy?



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

2360
1212
570

Large disagreement between experiments!

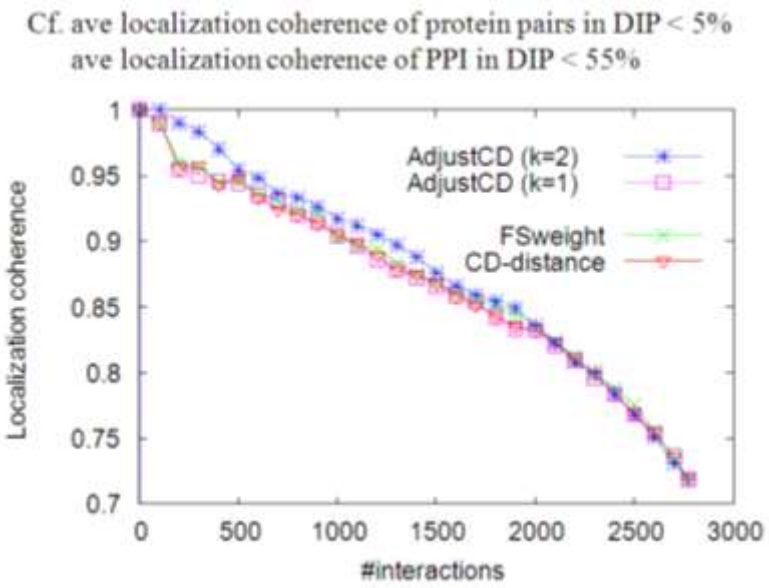
Chua & Wong. Increasing the Reliability of Protein Interactomes.
Drug Discovery Today, 13(15/16):652--658, 2008

Dealing with noise in PPIN

- **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



Czekanowski-Dice Distance

- **Functional distance between two proteins**

$$D(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}$$

- **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?

$\Rightarrow$ **Similarity can be defined as**

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

Brun, et al. *Genome Biology,* 5(1):R6, 2003

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**

$\Rightarrow$ **Rewriting this as**

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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

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

- **FS-Weight is slightly better in correlation w/ similarity for L1 & L2 neighbours**

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Reliability of Expt Sources

- **Diff expt sources have diff reliabilities**
  - Assign reliability to an interaction based on its expt sources
- **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**

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**

$\Rightarrow$ **Rewriting**

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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

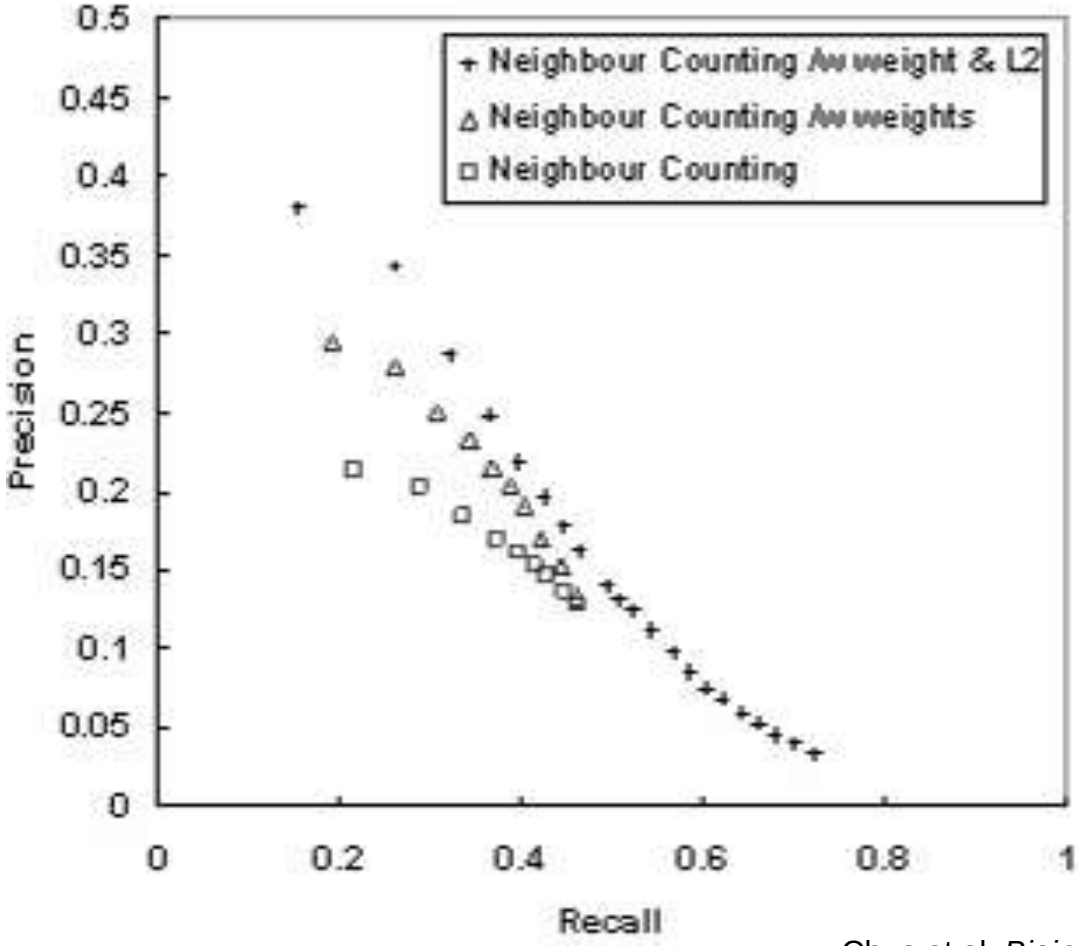
Integrating Reliability

- **FS-Weight 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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

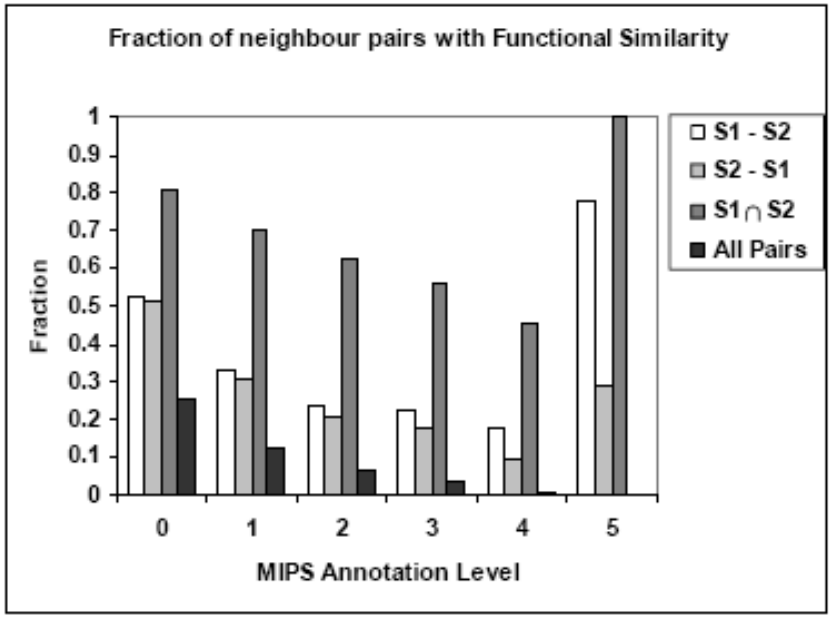
Improvement to Prediction Power by Majority Voting

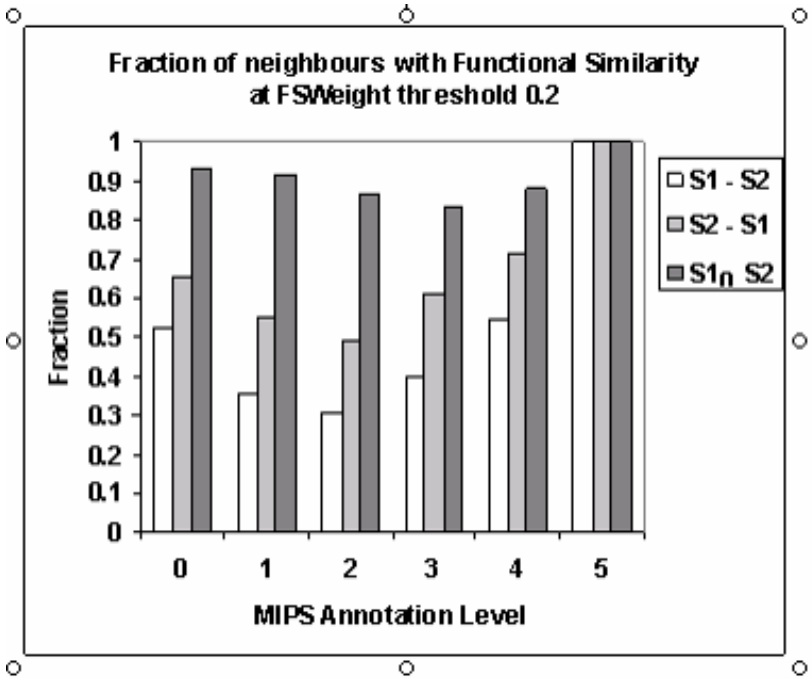


Considering only neighbours w/ FS weight > 0.2

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Improvement to Over-Rep of Functions in Neighbours





Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Use L1 & L2 Neighbours for Prediction

- **FS-weighted Averaging (FWA)**

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

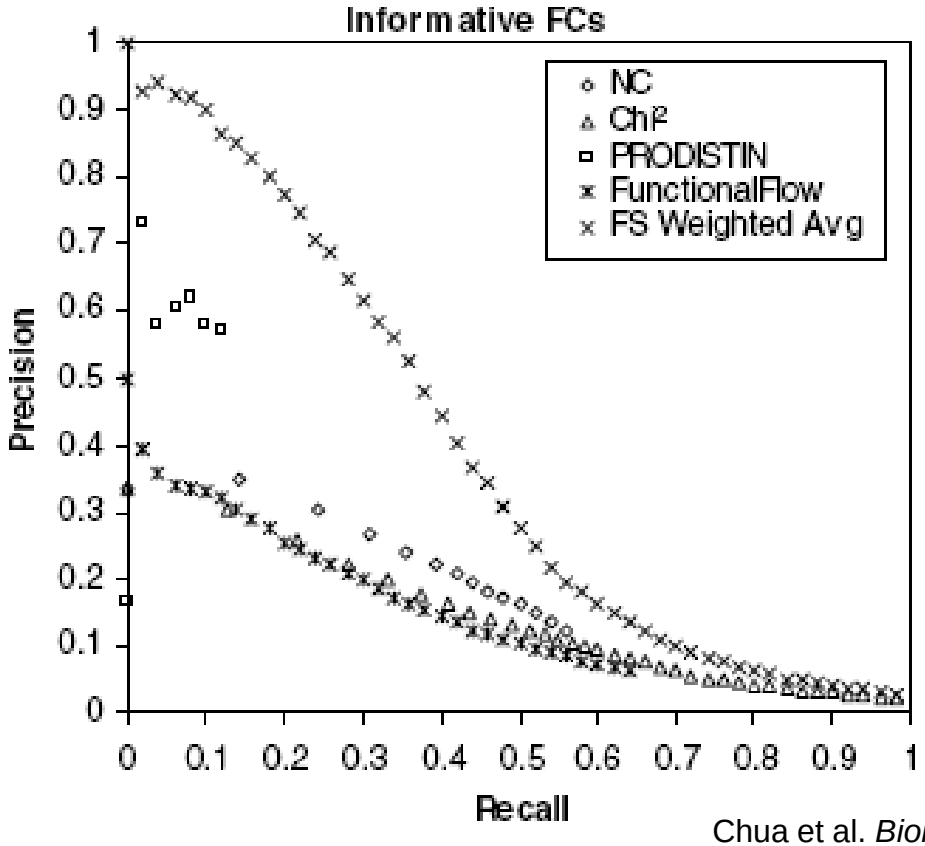
  - **r_{int} is fraction of all interaction pairs sharing function**
  - **λ is weight of contribution of background freq**
  - **δ(k, x) = 1 if k has function x, 0 otherwise**
  - **N_k is the set of interacting partners of k**
  - **π_x is freq of function x in the dataset**
  - **Z is sum of all weights**

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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Performance of FS-Weighted Averaging

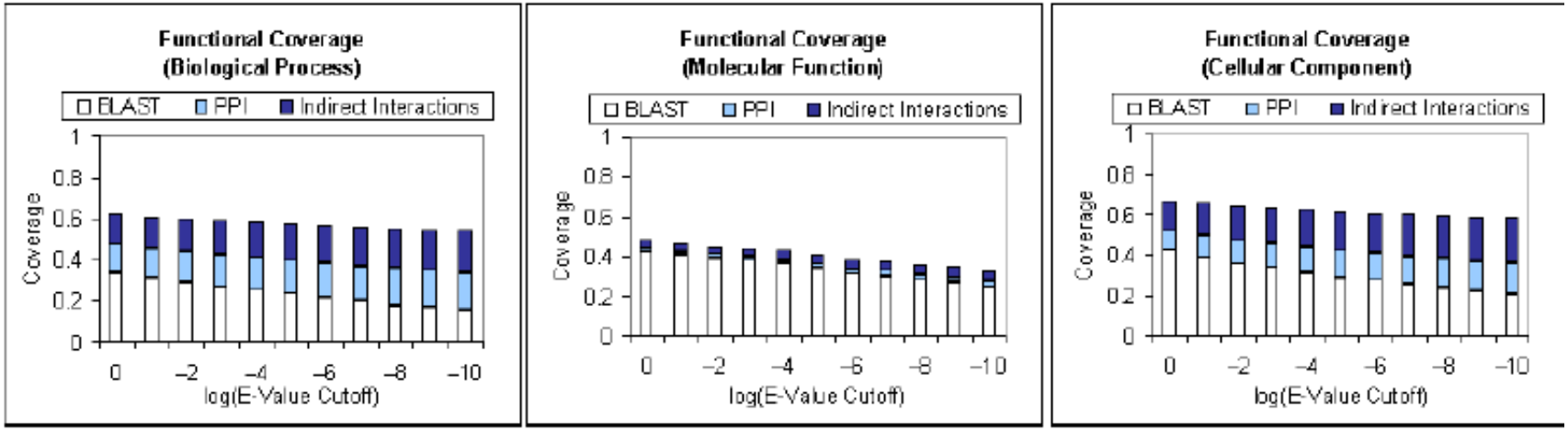
- **LOOCV comparison with Neighbour Counting, Chi-Square, PRODISTIN**



Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Freq of indirect functional association in other genomes

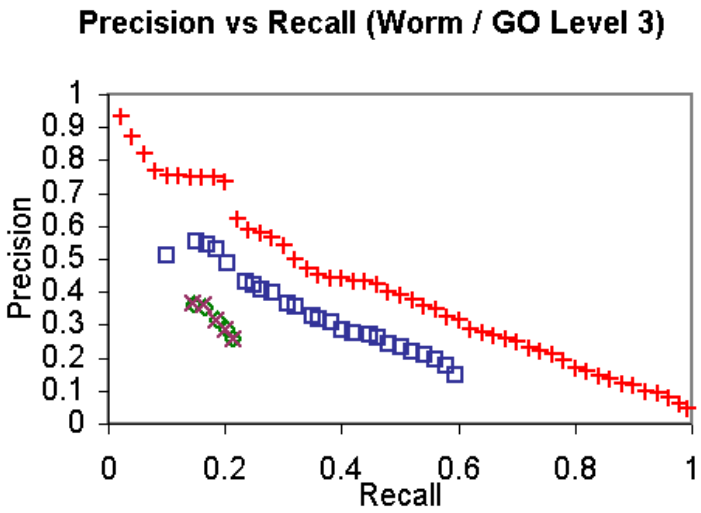
D. melanogaster

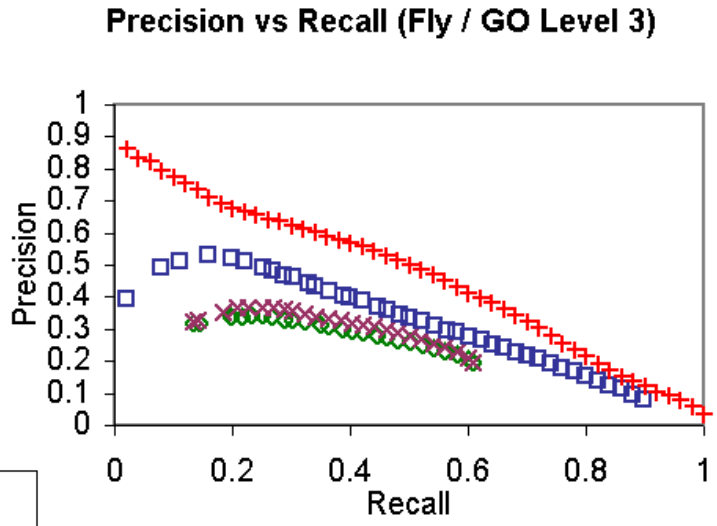


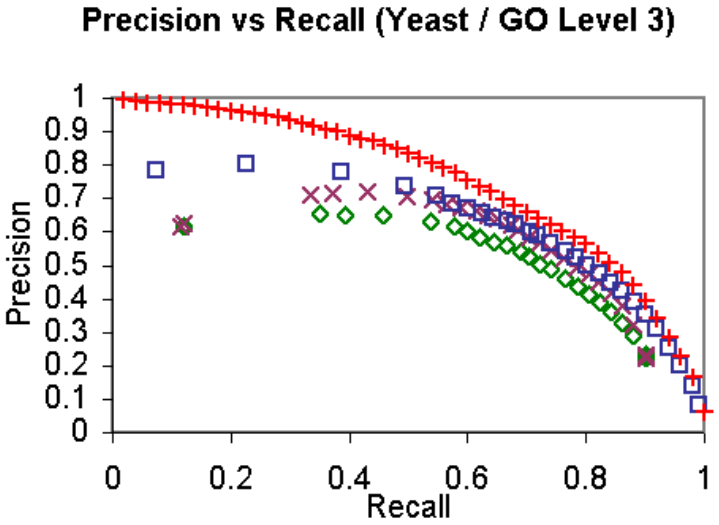
Genome	Annotation	S_1-S_2	S_2-S_1	$S_1 \cap S_2$	$S_1 \cup S_2$
S. cerevisiae	MIPS	0.007193	0.226574	0.463960	0.706872
D. melanogaster	GO	0.008801	0.168622	0.138138	0.315561
C. elegans	GO	0.007193	0.051237	0.061080	0.119510

Chua et al. Using Indirect Protein Interactions for the Prediction of Gene Ontology Functions. *BMC Bioinformatics*, 8(Suppl 4):S8, 2007

Effectiveness of FSWeighted Averaging in other genomes







- ◇ Neighbour Counting
- × NC (Weighted)
- □ NC (Weighted + L2)
- + Weighted Avg

Chua et al. Using Indirect Protein Interactions for the Prediction of Gene Ontology Functions. *BMC Bioinformatics*, 8(Suppl 4):S8, 2007

What have we learned?

- **Proteins with similar function are topologically close in PPIN**

⇒ **Assign protein to a function that is over represented in its neighborhood**

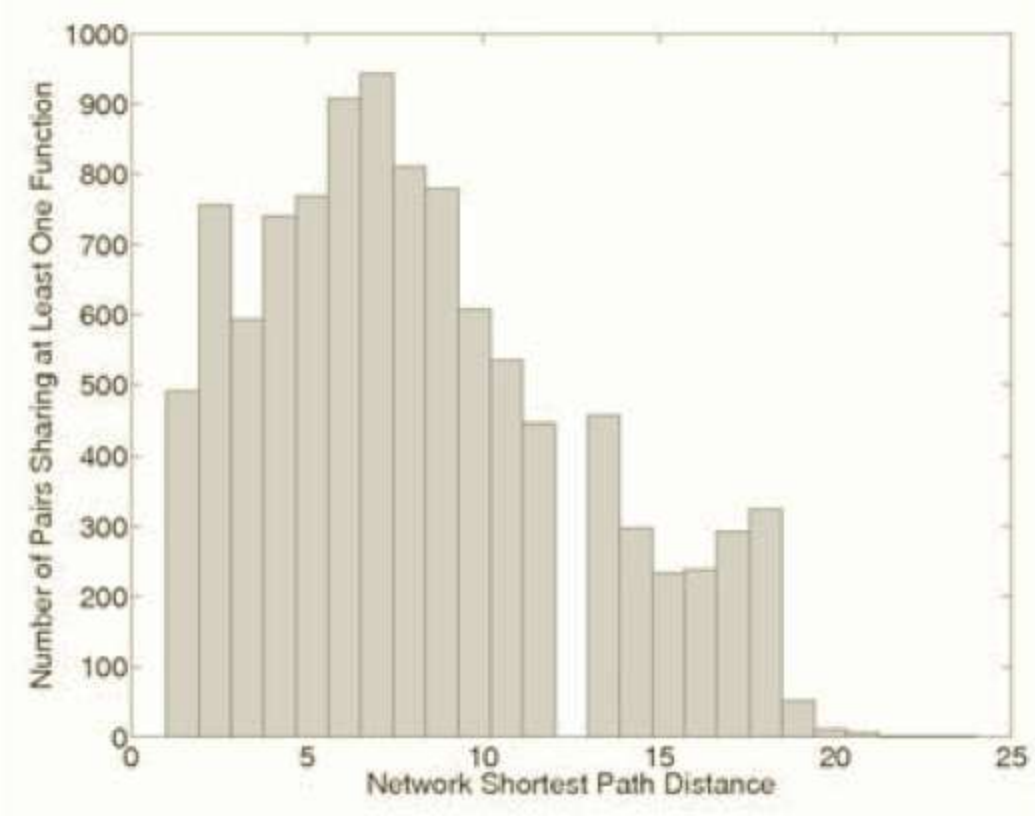
  – Indirect neighbors are useful

- **PPIN is noisy**

  – Not are neighbors are "real"

⇒ **Need to clean up the PPIN before "voting"**

But genes sharing annotations do not always interact…

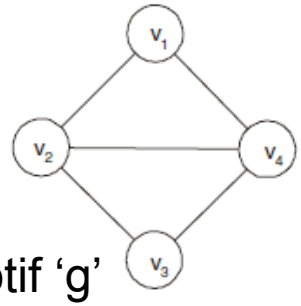


- **Similar functions are sometimes at large network distances**

Source: Bogdanov & Singh. *TCBB,* 7:208–217, 2010

Labeled Motifs

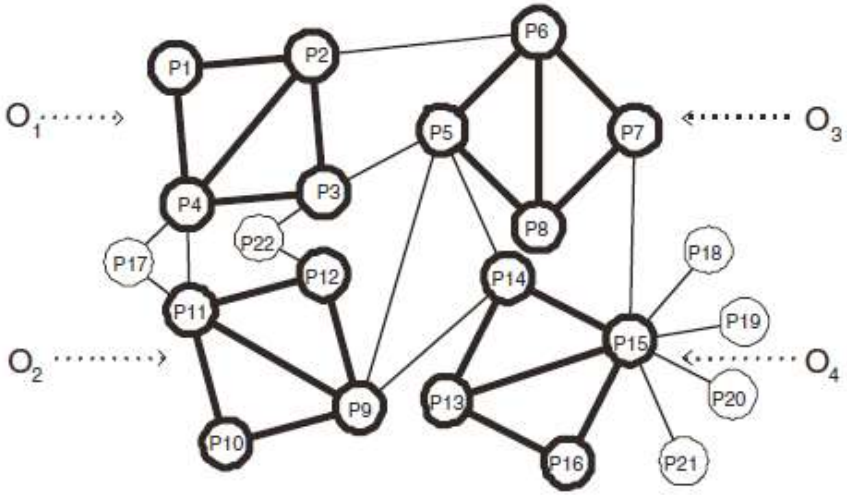
- **Proteins with similar function have interaction neighborhoods that are similar**



Network motif ‘g’

g

- **Assign a protein a function based on “network motif” that its neighborhood matches**

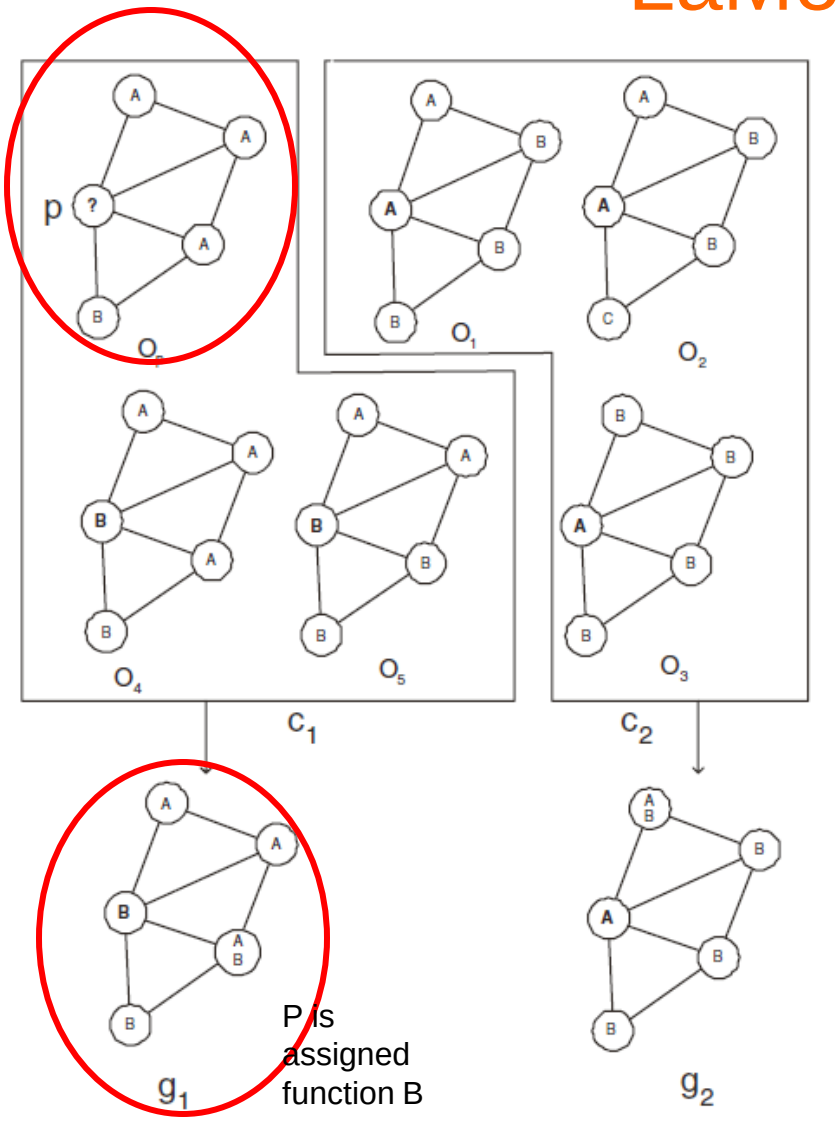


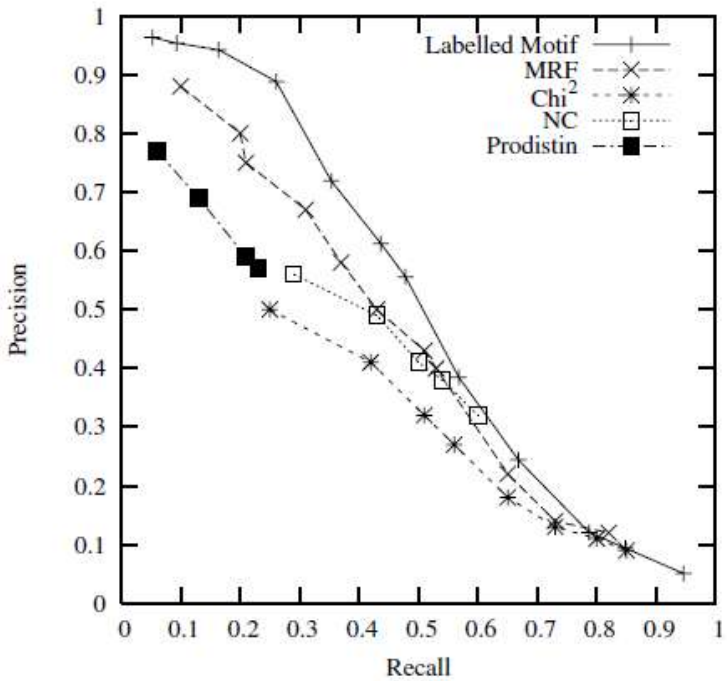
4 occurrences of ‘g’ in this PPIN

Image credit: Chen et al. *ICDE2007*, pp. *546–555*

Chen et al. *ICDE2007*, pp. *546–555*

LaMoFinder





- **Shortcoming**
  - Works only for proteins in subnets that can be mapped to network motifs

Pattern-Based Annotation Prediction (PAP)

- Kirac & Ozsoyoglu, *RECOMB2008*, pp 197-213
- **Find the best pairwise graph alignment of the functionally labeled subgraph rooted at the unknown protein to functionally labeled subgraphs rooted at other nodes in the protein interaction network**

- **Shortcoming**
  - Rely on topological matching of subnetworks

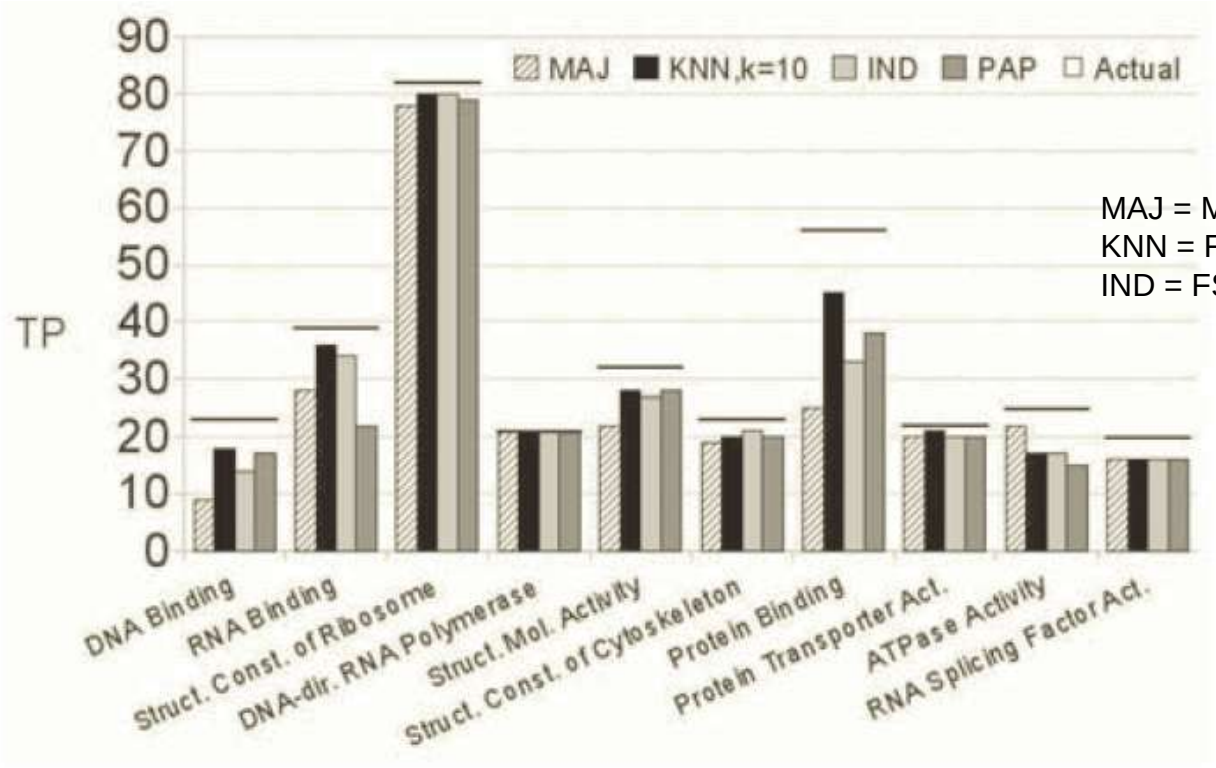
  ⇒Sensitive to noise & missing edges in PPIN

Functional Neighborhood Features

- Bogdanov & Singh. *TCBB,* 7:208–217, 2010
- **Predict function of an unknown protein v by weighted voting of the k proteins having most similar functional profiles to v**

- **Affinity of protein u to protein v**
  - $P_{u,v}$ = Prob of random walks from u to v
- **Affinity of protein v to function a**
  - $Sf_v(a) = \Sigma P_{u,v}$, over all proteins u having function a
- **Functional profile of a protein v**
  - $[Sf_v(a_1), \ldots, Sf_v(a_k)]$, normalized

Comparisons



MAJ = Majority Voting
KNN = Functional Neighborhood Features
IND = FSWeight

Fig. 10. Number of TP per GO molecular function (*FYI*, $T = 20$). The top two functions are considered as predictions for each of the methods. The horizontal bars represent the total number of TPs for each GO term.

- **Functional neighborhood features is slightly better than FSWeight**

Bogdanov & Singh. *TCBB,* 7:208–217, 2010

What have we learned?

- **Proteins with similar function can be far apart**

- **If the functional neighborhood features of two proteins are similar, they may have similar function**

$\Rightarrow$ **Assign protein to a function based on network motif (and generalizations thereof) that it matches**

References

- Wong. **Using biological networks in protein function prediction and gene expression analysis.** *Internet Math*, 7(4):274-298, 2011
- [Majority Voting, $\chi 2$] Hishigaki et al. **Assessment of prediction accuracy of protein function from protein-protein interaction data**. *Yeast*, 18:523-531, 2001
- [FSWeight] Chua et al. **Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions**. *Bioinformatics*, 22:1623-1630, 2006
- [LaMoFinder] Chen et al. **Labeling Network Motifs in Protein Interactomes for Protein Function Prediction.** *ICDE2007, 546–555*
- [PAP] Kirac & Ozsoyoglu. **Protein Function Prediction based on Patterns in Biological Networks.** *RECOMB2008*, 197–213
- [Functional Neighborhood Features] Bogdanov & Singh. **Molecular Function Prediction Using Neighborhood Features.** *TCBB,* 7:208–217, 2010

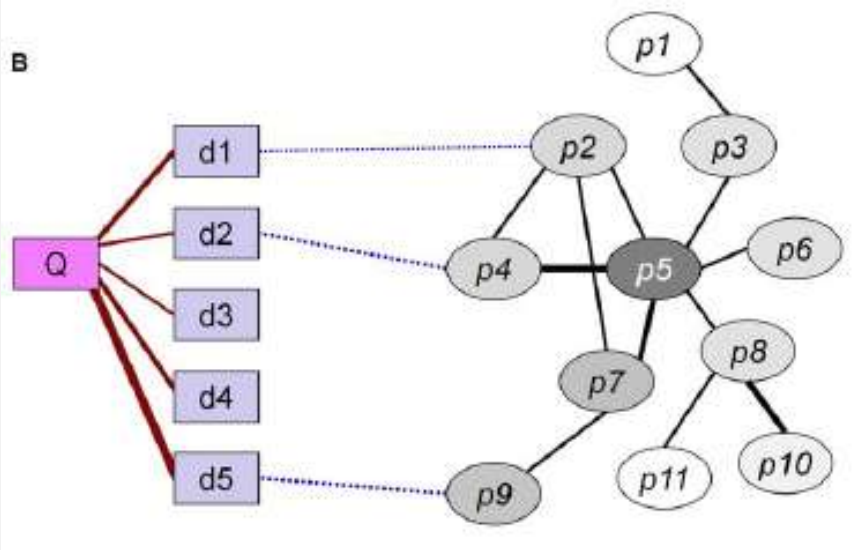
Using Biological Networks, Part 4: *Other Applications*

Limsoon Wong

NUS
National University of Singapore

Part 4: Other applications of biological networks

- **Epistatic interaction mining**
- **Disease causal gene prioritization**
- **Protein complex prediction**



Emily et al. *European Journal of Human Genetics*, 17(10):1231-1240, 2009.

Epistatic Interaction Mining

- **GWAS have linked many SNPs to diseases, but many genetic risk factors still unaccounted for**
- **Proteins coded by genes interact in cell**

⇒ **Some SNPs affect the phenotype in combination with other SNPs; i.e., epistasis**

- **Exhaustive search for epistatic effects has to test many combinations (>100,000^2) of SNPs**
  - Hard to get statistical significance
  - Take long time to run on computers

⇒ **Use biological networks to narrow the search for two-locus epistasis**

Vanunu et al. *PLoS Computational Biology*, 6(1):e1000641, 2010

Disease Causal Gene Prioritization

- **Genes causing the same or similar diseases tend to lie close to one another in PPIN**

- **Given disease Q. Look in PPIN for proteins that interact with many causal genes of diseases similar to Q**

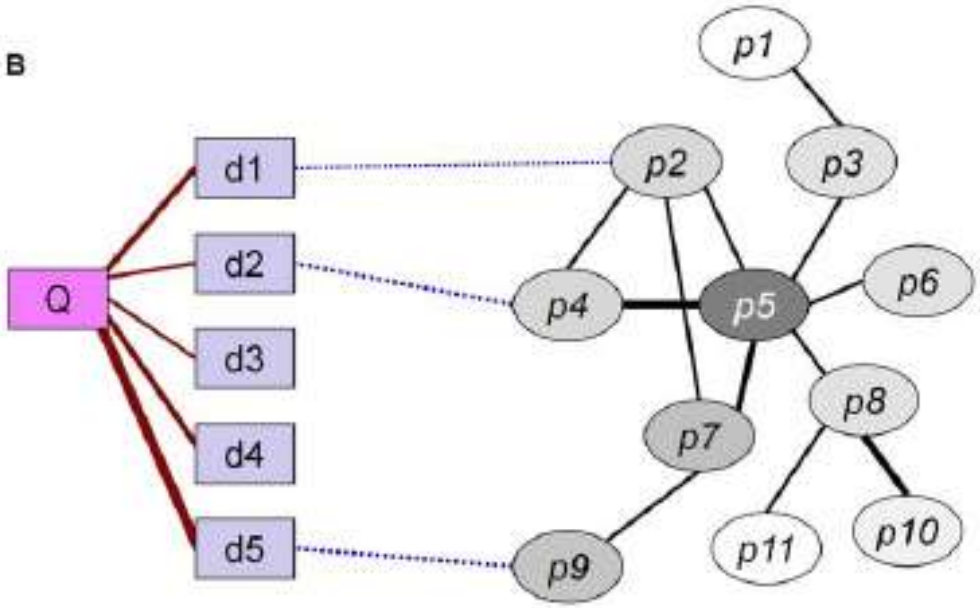
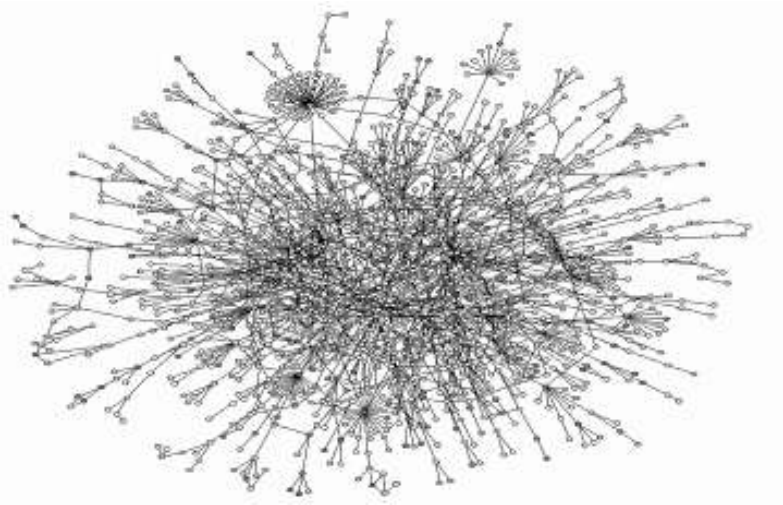


Figure 1. Illustration of the PRINCE algorithm. A query disease, denoted *Q*, has varying degrees of phenotypic similarity with other diseases, denoted *d1–d5* (marked with maroon lines, where thicker lines represent higher similarity). Known causal genes for these similar diseases are connected by dashed blue lines and used as the prior information. *p1–p11* comprise the protein set of a protein-protein interaction network, where interactions are marked with black lines and thicker lines denote edges with higher confidence. A scoring function that is smooth over the network is computed using an iterative network propagation method. At every iteration of the algorithm, each protein pumps flow to its neighbors and receives flow from them. Protein colors correspond to the flow they receive in a specific iteration, the darker the color the higher the flow. (A):

Liu et al. *Bioinformatics*, 25(15):1891-1897, 2009

Protein Complex Prediction

- **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 &time; viz., a protein complex

- **Many complexes have highly connected cores in PPIN ➔ Find complexes by clustering**
- **Issue: How to identify low edge density complexes?**

References

- Emily et al. **Using biological networks to search for interacting loci in genome-wide association studies.** *European Journal of Human Genetics*, 17(10):1231-1240, 2009

- Vanunu et al. **Associating genes and protein complexes with disease via network propagation.** *PLoS Computational Biology*, 6(1):e1000641, 2010

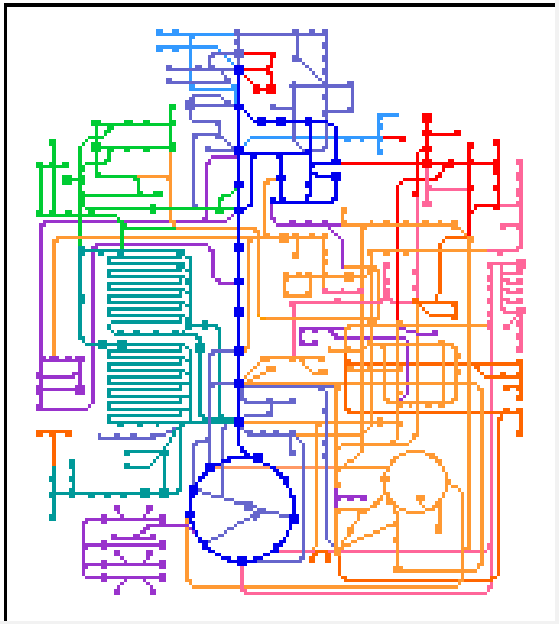
- Liu et al. **Complex Discovery from Weighted PPI Networks.** *Bioinformatics*, 25(15):1891-1897, 2009

Issues in Using Biological Networks

Limsoon Wong

NUS
National University
of Singapore

How good are available sources of pathway & PPI Network?



- **Sources of pathway & PPIN**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**
- **PPIN prediction**

Sources of Protein Interactions

Database	# nodes, # edges	URL	Build Focus	Reference
BioGRID	10k, 40k	http://thebiogrid.org	Literature	(Stark *et al.*, 2006)
DIP	2.6k, 3.3k	http://dip.doe-mbi.ucla.edu	Literature	(Xenarios *et al.*, 2002)
HPRD	30k, 40k	http://www.hprd.org	Literature	(Prasad *et al.*, 2009)
IntAct	56k, 267k	http://www.ebi.ac.uk/intact	Literature	(Aranda *et al.*, 2010)
MINT	30k, 90k	http://mint.bio.uniroma2.it/mint	Literature	(Chatr-aryamontri *et al.*, 2007)
STRING	5200k, ?	http://string-db.org	Literature, Prediction	(Szklarczyk *et al.*, 2011)

and Protein Complexes

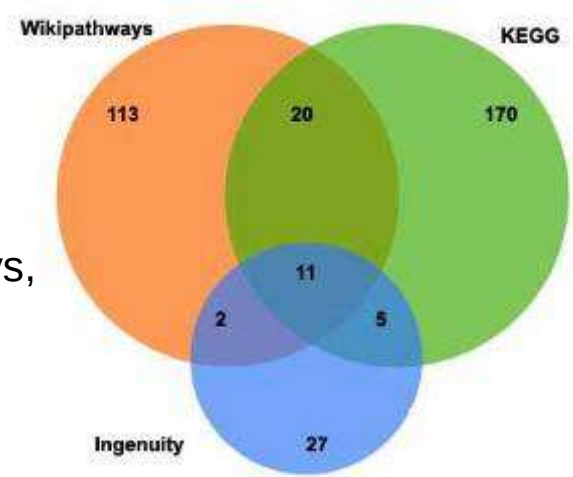
- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, *NAR*, 2010

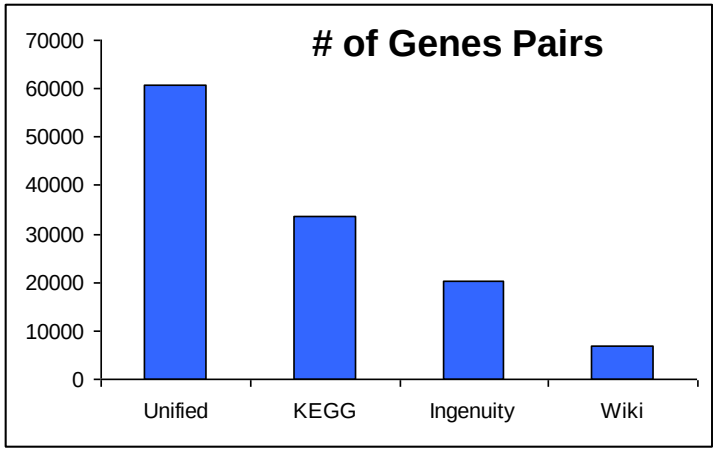
Sources of Biological Pathways

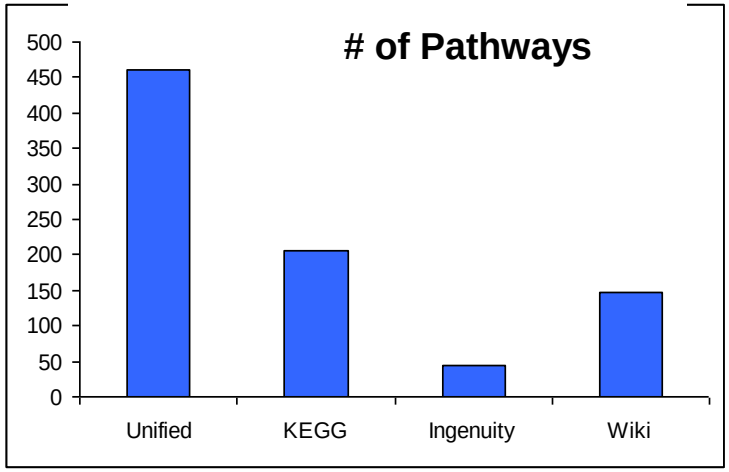
Database	Remarks
KEGG	KEGG (`http://www.genome.jp/kegg`) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.
WikiPathways	WikiPathways (`http://www.wikipathways.org`) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.
Reactome	Reactome (`http:://www.reactome.org`) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.
Pathway Commons	Pathway Commons (`http://www.pathwaycommons.com`) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.
PathwayAPI	PathwayAPI (`http://www.pathwayapi.com`) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh *et al.*, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.

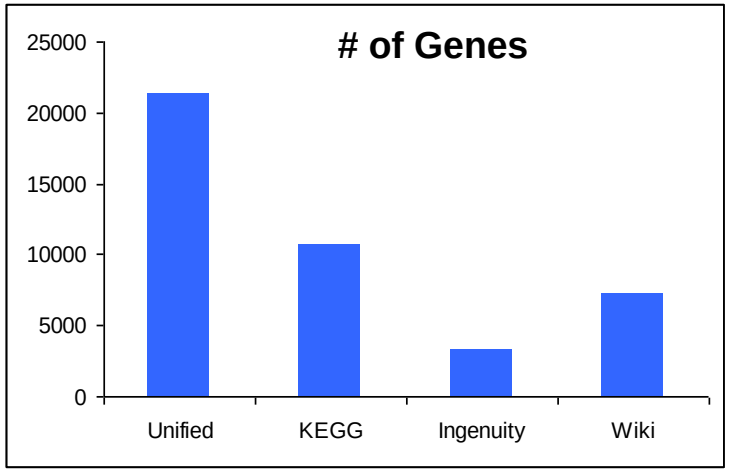
Low Comprehensiveness of Pathway Sources

Human pathways in Wikipathways, KEGG, & Ingenuity





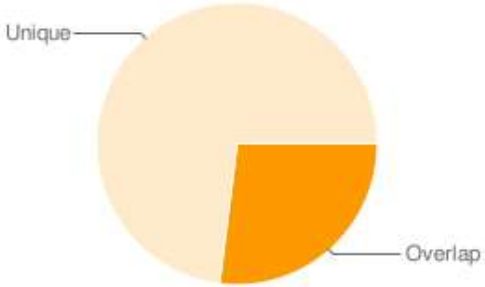




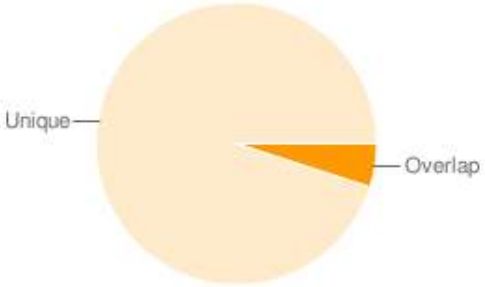
Soh et al. Consistency, Comprehensiveness, and Compatibility of Pathway Databases. *BMC Bioinformatics*, 11:449, 2010.

Low Consistency of Pathway Sources

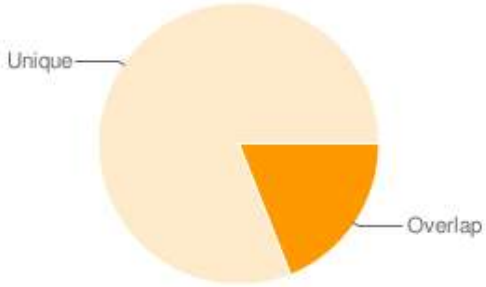
Gene Pair Overlap



Wiki vs KEGG

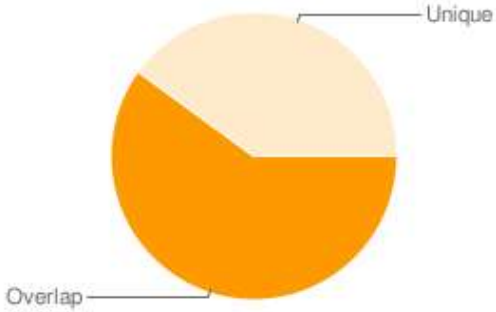


Wiki vs Ingenuity

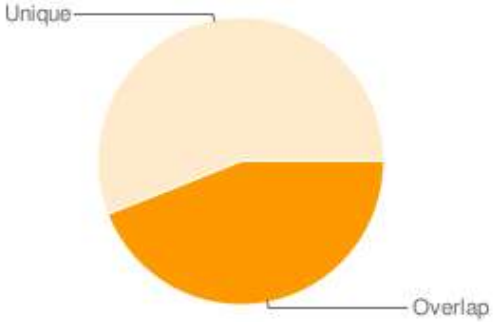


KEGG vs Ingenuity

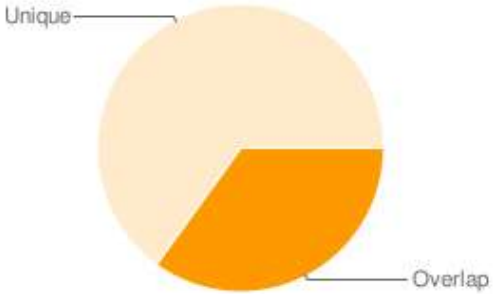
Gene Overlap



Wiki vs KEGG



Wiki vs Ingenuity



KEGG vs Ingenuity

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

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%

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

Pathway sources are curated. They are incomplete; but they have few errors. ➔ Makes sense to combine them. But…

Incompatibility Issues

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

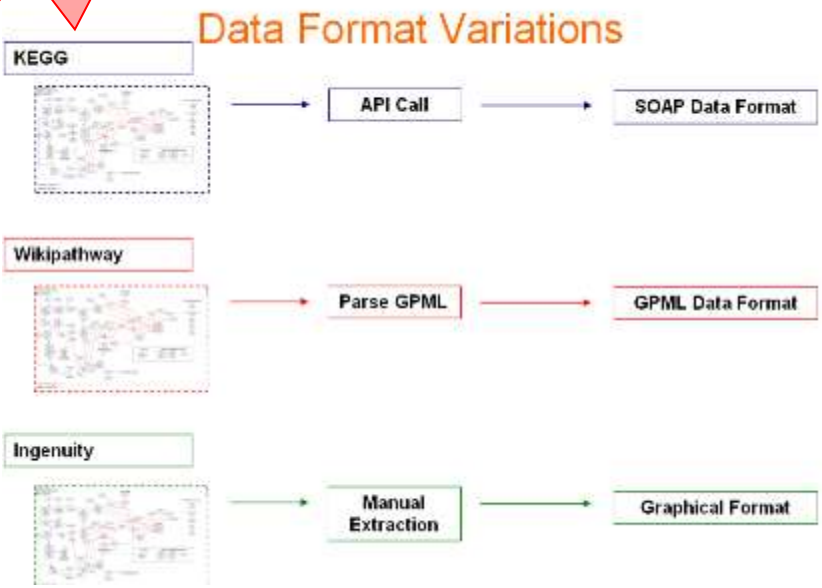


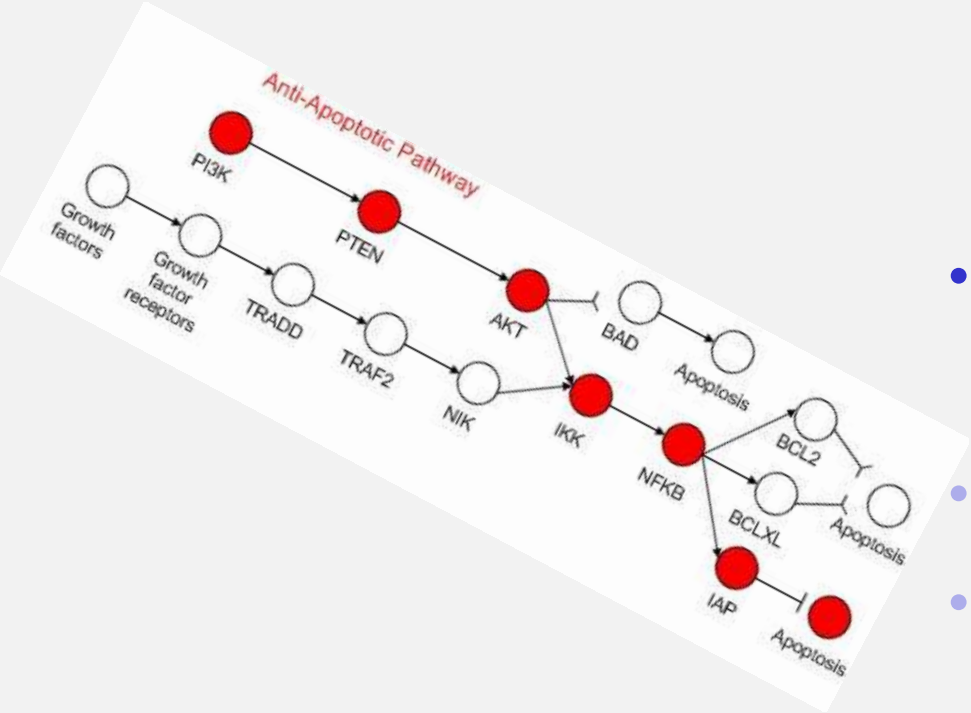
Image credit: Donny Soh's PhD dissertation, 2009

The preceding analyses hide an intricate issue…

The same pathways in the different sources are often given different names.

So how do we even know two pathways are the same and should be compared / merged?

How good are available sources of pathway information?



- **Sources of pathway info**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**
- **PPIN prediction**

Possible Ways to Match Pathways

- **Match based on name (LCS)**
  - Pathways w/ similar name should be the same pathway
  - But annotations are very noisy

  ⇒Likely to mismatch pathways?

  ⇒Likely to match too many pathways?

- **Are the followings good alternative approaches?**
  - Match based on overlap of genes
  - Match based on overlap of gene pairs

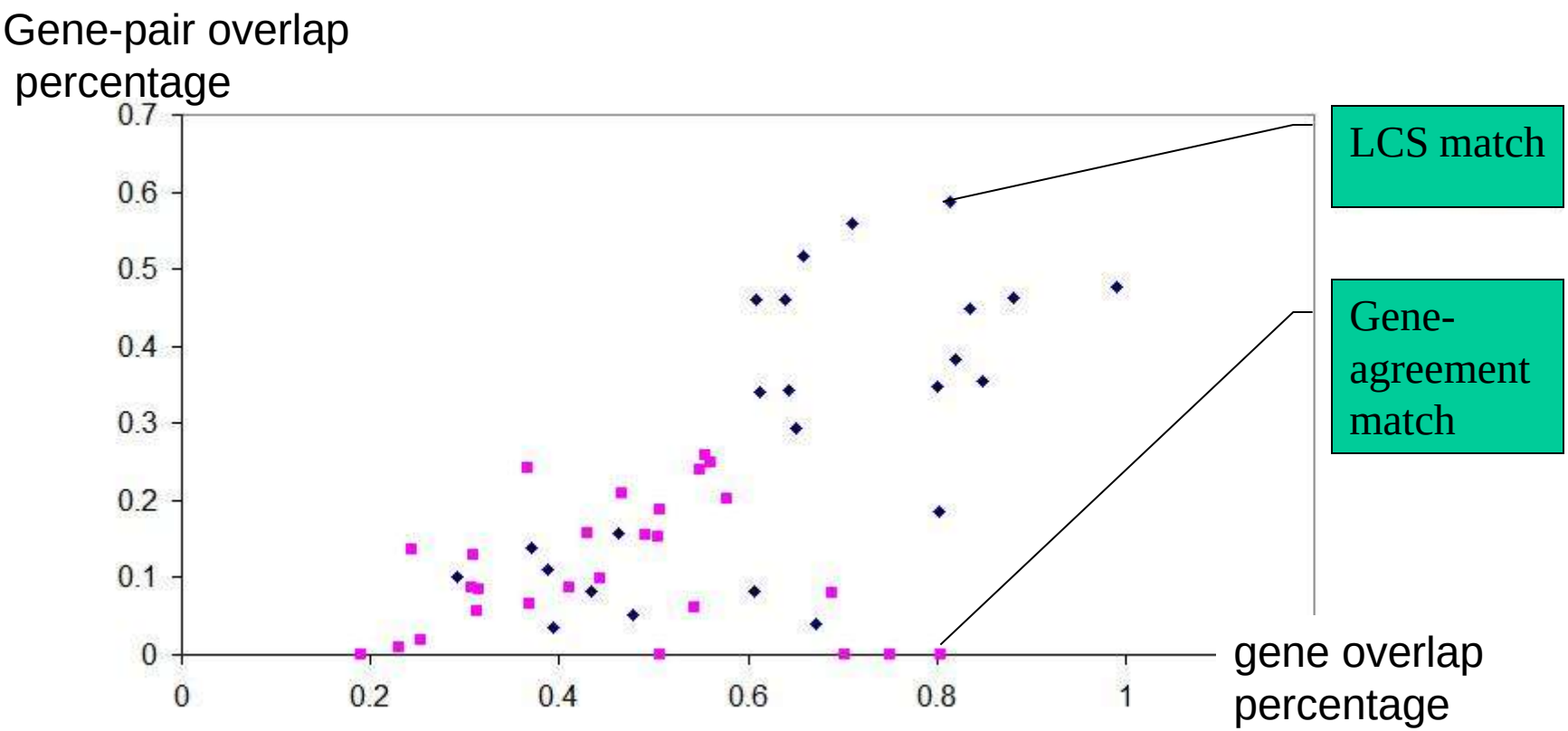
LCS vs Gene-Agreement Matching

- **Accuracy**
  - 94% of LCS matches are in top 3 gene agreement matches
  - 6% of LCS matches not in top 3 of gene agreement matches; but their gene-pair agreement levels are higher
- **Completeness**
  - Let Pi be pathway in db A that LCS cannot find match in db B
  - Let Qi be pathway in db B with highest gene agreement to Pi
  - Gene-pair agreement of Pi-Qi is much lower than pathway pairs matched by LCS

LCS is better than gene-agreement based matching!

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

LCS vs Gene-Agreement Matching

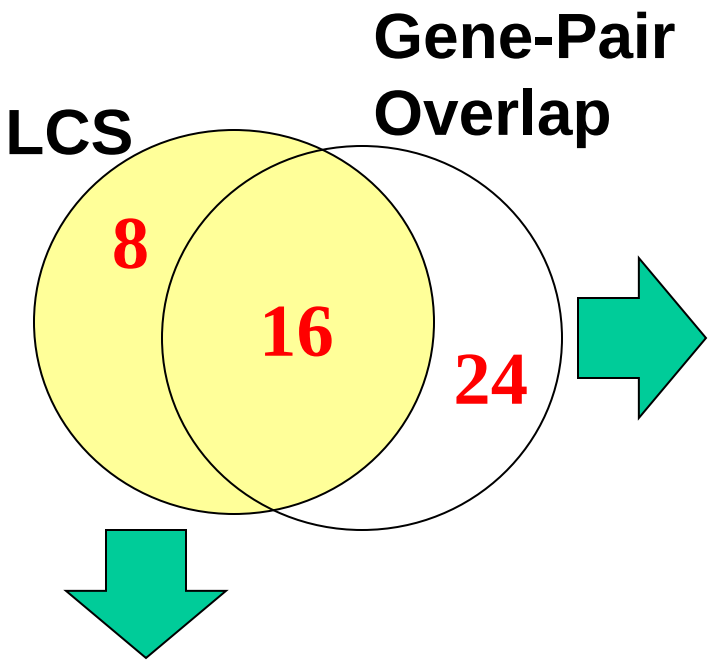


- **LCS consistently has higher gene-pair agreement**

⇒ LCS is better than gene-agreement based matching!

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

LCS vs Gene-Pair Agreement Matching



ErbB signaling pathway	JAK/Stat Signaling
Calcium signaling pathway	Synaptic Long Term Potentiation
Apoptosis	Toll-like receptor signaling pathway
VEGF signaling pathway	Axonal Guidance Signaling
Gap junction	PPAR-alpha/RXR-alpha Signaling
Natural killer cell mediated cytotoxicity	Fc Epsilon RI Signaling
T cell receptor signaling pathway	Axonal Guidance Signaling
B cell receptor signaling pathway	Axonal Guidance Signaling
Olfactory transduction	cAMP-mediated Signaling
GnRH signaling pathway	B Cell Receptor Signaling
Melanogenesis	Wnt Signaling Pathway and Pluripotency
Type II diabetes mellitus	Insulin Recpetor Signaling
Colorectal cancer	Toll-like receptor signaling pathway
Renal cell carcinoma	Axonal Guidance Signaling
Pancreatic cancer	PTEN Signaling
Endometrial cancer	PTEN Signaling
Glioma	ERK/MAPK Signaling
Prostate cancer	JAK/Stat Signaling
Basal cell carcinoma	Wnt Signaling Pathway and Pluripotency
Melanoma	FGF Signaling
Chronic myeloid leukemia	GM-CSF Signaling
Acute myeloid leukemia	PTEN Signaling
Small cell lung cancer	Toll-like receptor signaling pathway
Non-small cell lung cancer	GM-CSF Signaling

The 24 pathway pairs singled out by maximal gene-pair overlap

Regulation of actin cytoskeleton	Regulation of Actin Cytoskeleton
Wnt signaling pathway	Wnt Signaling Pathway
T cell receptor signaling	t cell receptor Signaling
VEGF signaling	VEGF Signaling
MAPK signaling	MAPK Cascade
Apoptosis	Apoptosis
Apoptosis	Apoptosis Signaling
Toll-like receptor	Toll-like receptor signaling pathway

The 8 pathway pairs singled out by LCS

Note: We consider only pathway pairs that have at least 20 reaction overlap.

LCS vs Gene-Pair Agreement Matching

- **Gene-pair agreement match will miss when**
  - Pathway P in db A has few overlap with pathway P in db B due to incompleteness of db, even if pathway name matches perfectly!
  - Example: wnt signaling pathway, VEGF signaling pathway, MAPK signaling pathway, etc. in KEGG don't have largest gene-pair overlap w/ corresponding pathways in Wikipathways & Ingenuity

$\Rightarrow$ **Bad for getting a more complete unified pathway P**

LCS vs Gene-Pair Agreement Matching

- **Pathways having large gene-pair overlap are not necessarily the same pathways**

- **Examples**
  - "Synaptic Long Term Potentiation" in Ingenuity vs "calcium signalling" in KEGG
  - "PPAR-alpha/RXR-alpha Signaling" in Ingenuity vs "TGF-beta signaling pathway" in KEGG

$\Rightarrow$ **Difficult to set correct gene-pair overlap threshold to balance against false positive matches**

… so we match pathways by LCS

- **Having found a good way to match up pathways in different datasources, we proceeded to build a big unified pathway db**

PathwayAPI
= KEGG
+ Wikipathways
+ Ingenuity

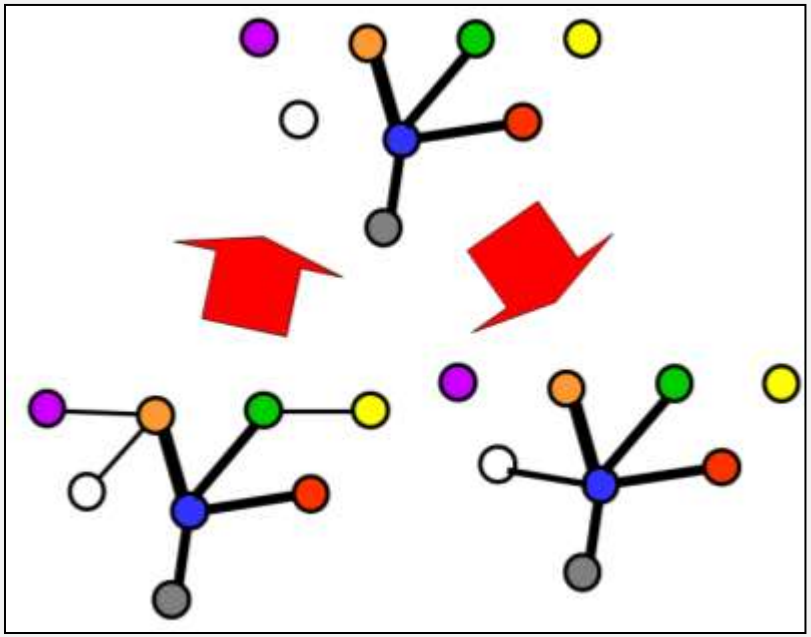
Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. Consistency, Comprehensiveness, and Compatibility of Pathway Databases. *BMC Bioinformatics*, 11:449, 2010.

What have we learned?

- **Significant lack of concordance betw db's**
  - Level of consistency for genes is 0% to 88%
  - Level of consistency for genes pairs is 0%-61%
  - Most db contains less than half of the pathways in other db's

- **Matching pathways by name is better than matching by gene overlap or gene-pair overlap**

How good are available sources of pathway & PPI Network?

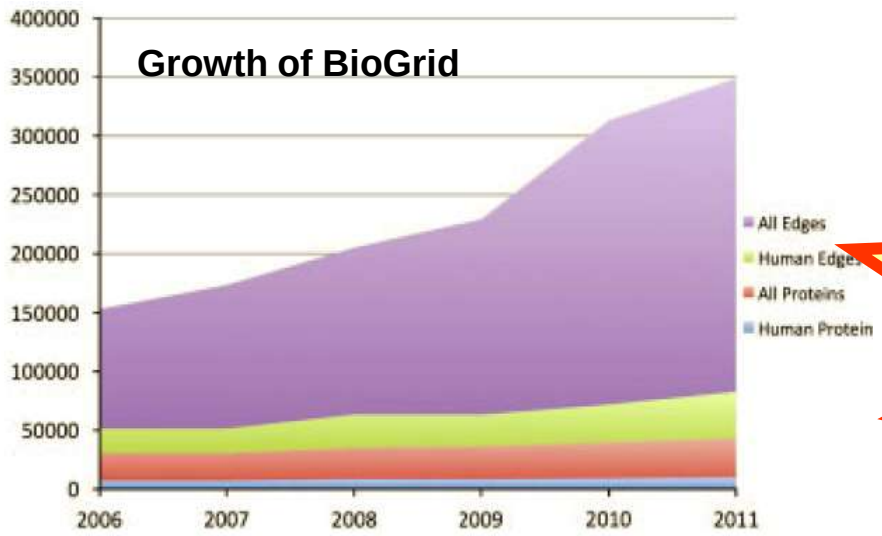


- **Sources of pathway & PPIN**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**
- **PPIN prediction**

PPI Detection Assays

- **Many high-throughput assays for PPIs**
  - Y2H
  - TAP
  - Synthetic lethality

Generating *large amounts* of expt data on PPIs can be done with ease



- **But …**

High-throughput approaches sacrifice quality for ***quantity***: (a) limited or biased coverage: ***false negatives***, & (b) high error rates: ***false positives***

Noise in PPI Networks

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

- **High level of noise**

$\Rightarrow$ **Need to clean up before making inference on PPI networks**

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

Dealing with noise in PPIN

- **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

Brun, et al. *Genome Biology,* 5(1):R6, 2003

Czekanowski-Dice Distance

- **Given a pair of proteins (u, v) in a PPI network**
  - N_u = the set of neighbors of u
  - N_v = the set of neighbors of v
- **CD(u,v) =** $$\frac{2\,|N_u \cap N_v|}{|N_u| + |N_v|}$$
- **Consider relative intersection size of the two neighbor sets, not absolute intersection size**
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, CD(u,v)=1
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, CD(u,v)=1

Liu et al. *GIW2008*, pp. 138-149

Iterated CD-Distance

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

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

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

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

A thought…

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

- **Weight of interaction reflects its reliability**

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

Liu et al. *GIW2008*, pp. 138-149

Iterated CD-Distance

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

Validation

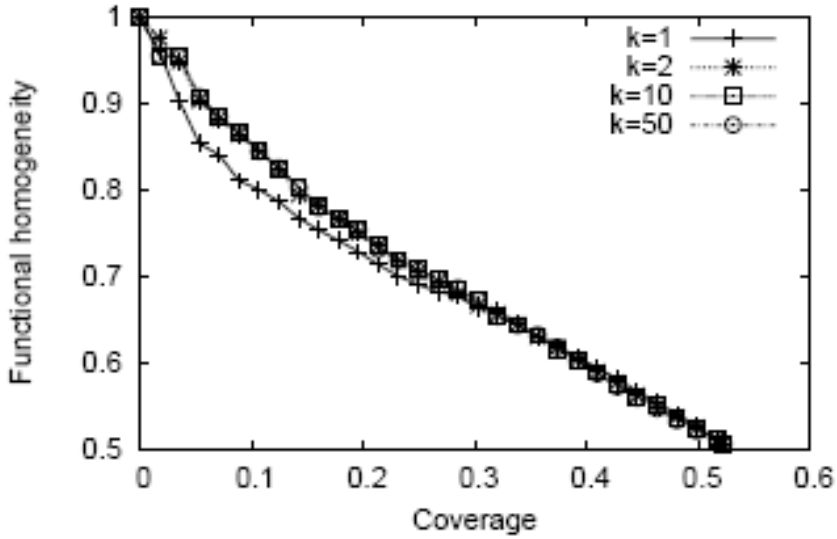
- **DIP yeast dataset**
  - Functional homogeneity is 32.6% for PPIs where both proteins have functional annotations and 3.4% over all possible PPIs
  - Localization coherence is 54.7% for PPIs where both proteins have localization annotations and 4.9% over all possible PPIs

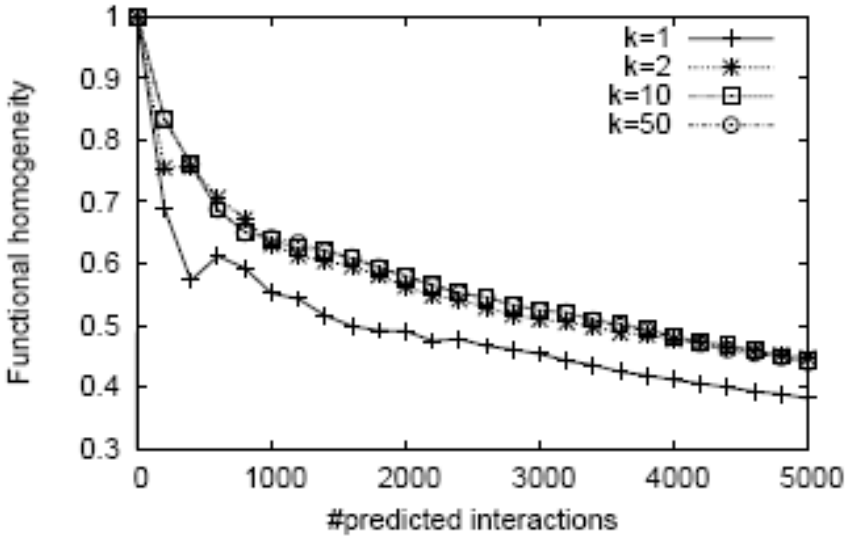
- **Let's see how much better iterated CD-distance is over the baseline above, as well as over the original CD-distance/FS-weight**

Liu et al. *GIW2008*, pp. 138-149

How many iteration is enough?

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





- **Iterated CD-distance achieves best performance wrt functional homogeneity at k=2**
- **Ditto wrt localization coherence (not shown)**

Liu et al. *GIW2008*, pp. 138-149

How many iteration is enough?

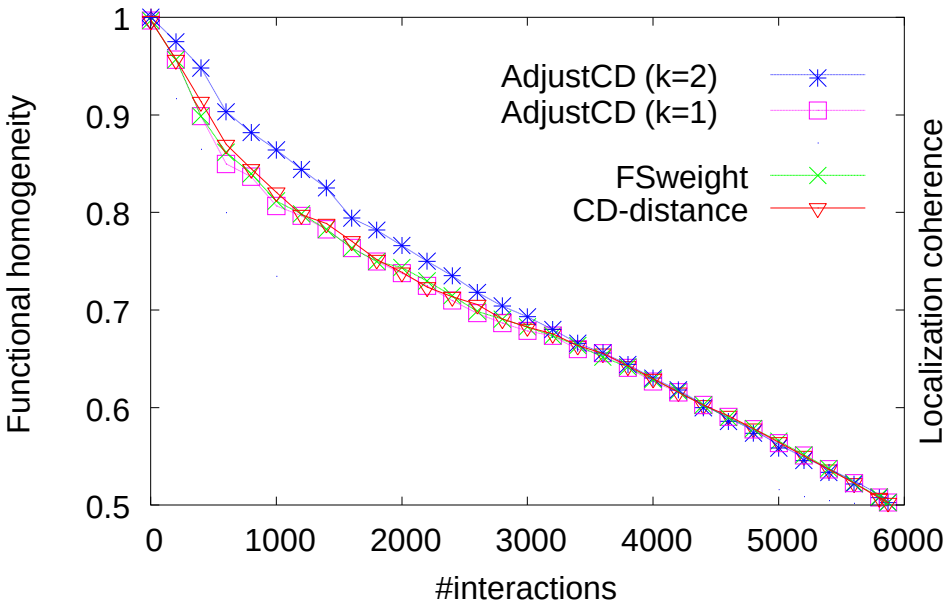
noise level	k	#common PPIs	avg_rank_diff	avg_score_diff
100%	1	5669	540.21	0.10
	2	5870	144.86	0.02
	20	5849	67.00	0.01
300%	1	5322	881.77	0.18
	2	5664	367.45	0.06
	20	5007	249.85	0.02
500%	1	5081	1013.14	0.23
	2	5502	625.46	0.12
	20	5008	317.33	0.05
1000%	k=1	4472	1187.10	0.28
	k=2	5101	1021.69	0.27
	k=20	5264	614.66	0.13

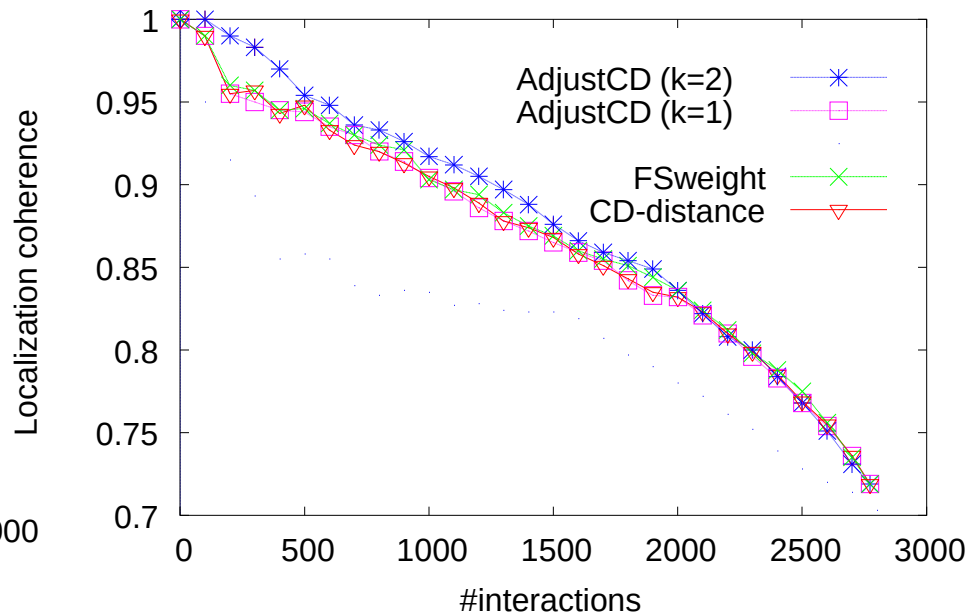
- **Iterative CD-distance at diff k values on noisy network**

⇒ # of iterations depends on amt of noise

Identifying False Positive PPIs

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



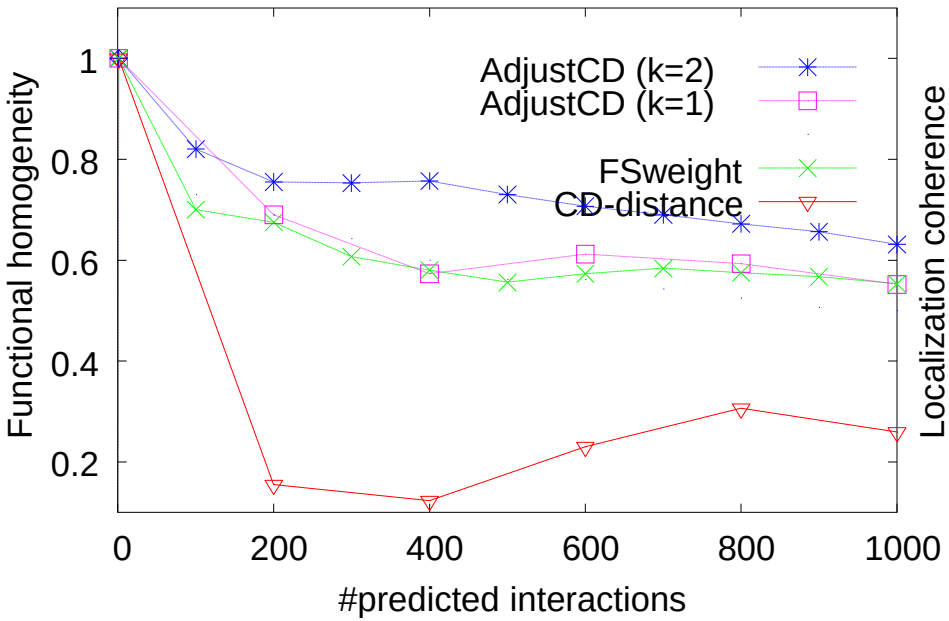


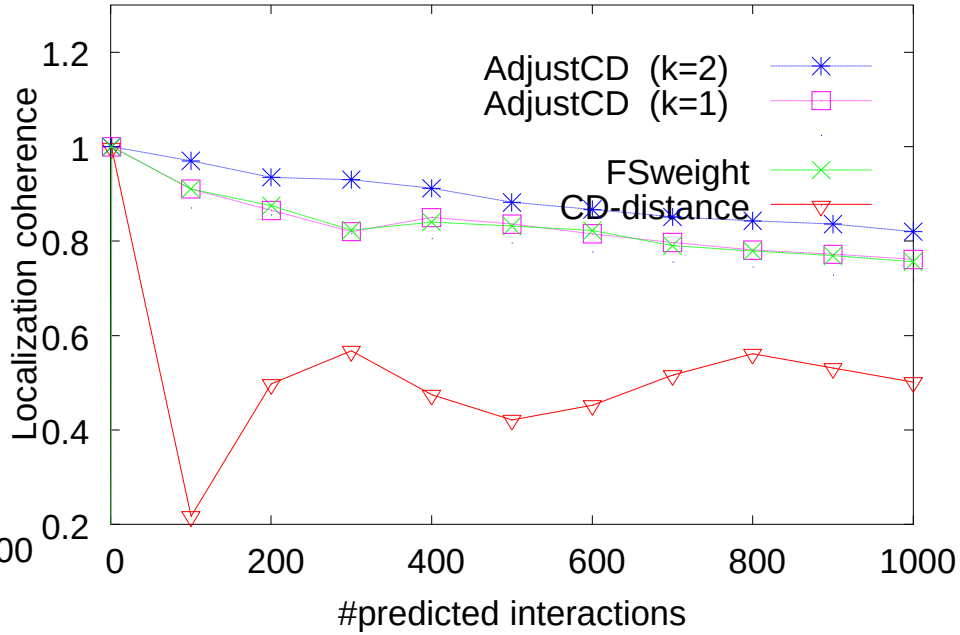
- **Iterated CD-distance is an improvement over previous measures for assessing PPI reliability**

Liu et al. *GIW2008*, pp. 138-149

Identifying False Negative PPIs

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





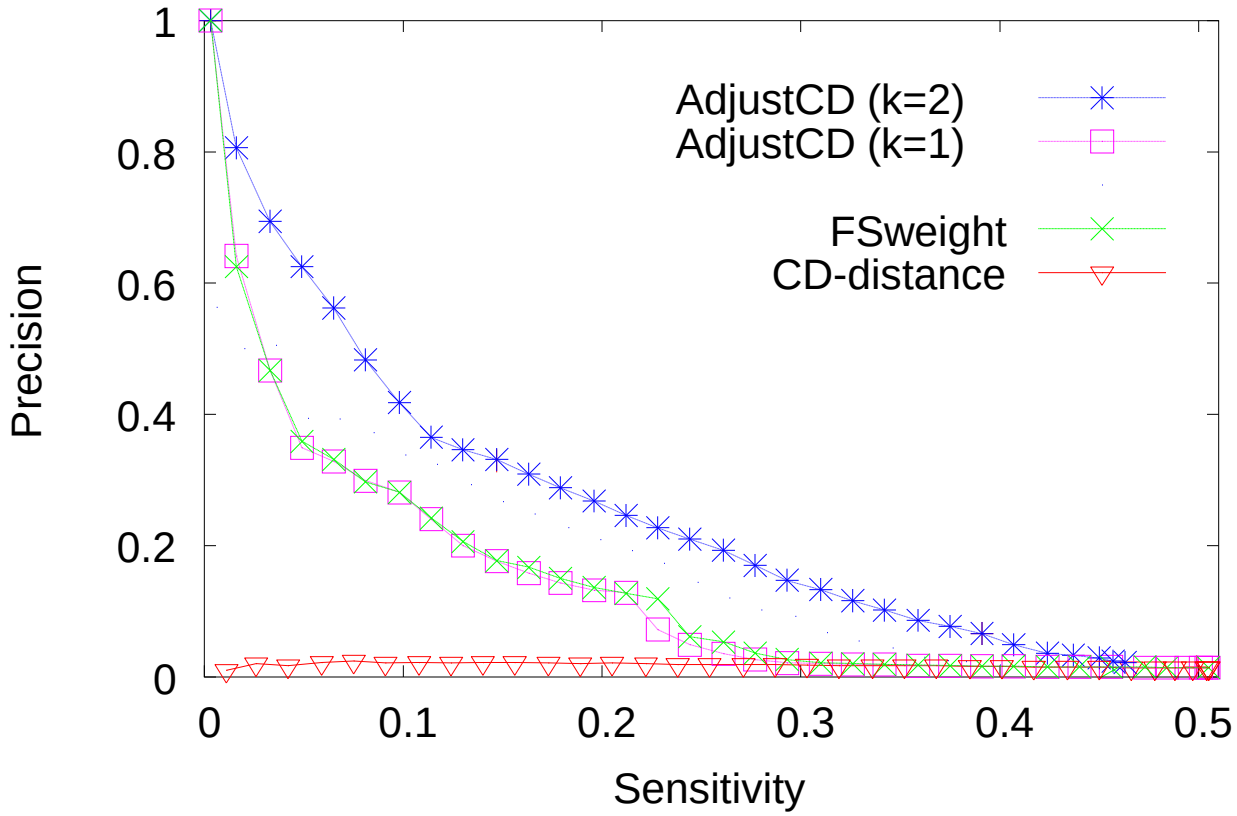
- **Iterated CD-distance is an improvement over previous measures for predicting new PPIs**

Liu et al. *GIW2008*, pp. 138-149

5-Fold Cross-Validation

- **DIP core dataset**
  - Ave # of proteins in 5 groups: 986
  - Ave # of interactions in 5 training datasets: 16723
  - Ave # of interactions in 5 testing datasets: 486591
  - Ave # of correct answer interactions: 307
- **Measures:**
  - sensitivity =TP/(TP + FN)
  - specificity =TN/(TN + FP)
    - **#negatives >> #positives, specificity is always high**
    - **>97.8% for all scoring methods**
  - precision =TP/(TP + FP)

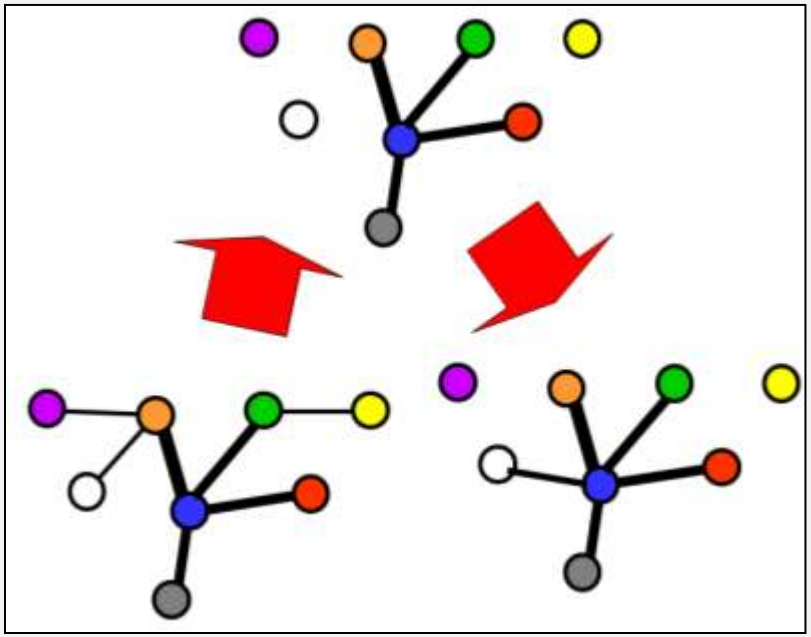
5-Fold X-Validation



- **Iterated CD-distance is an improvement over previous measures for identifying false positive & false negative PPIs**

Liu et al. *GIW2008*, pp. 138-149

How good are available sources of pathway & PPI Network?



- **Sources of pathway & PPIN**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**
- **PPIN prediction**

Shoemaker & Panchenko. *PLoS Computational Biology*, 3(4):e43, 2007

PPI Prediction Methods

Method Name	Protein/Domain Interaction	Physical Interaction/ Functional Association
Gene co-expression	P	F
Synthetic lethality	P	F
Gene cluster and gene neighbor	P	F
Phylogenetic profile	P, D	F
Rosetta Stone	P	F
Sequence co-evolution	P, D	F
Classification	P, D	P
Integrative	P, D	P
Domain association	D	P
Bayesian networks	P, D	F, P
Domain pair exclusion	D	P
p-Value	D	P

Second column shows if method is designed to predict protein (P) or domain (D) interactions (note that predicted domains can also be used for verifying protein interactions).
Third column shows if the method can be used to infer direct physical interaction (P) or indirect functional association (F).

You can also use our earlier topology scores, e.g, CD-distance to predict novel PPIs

PPI Prediction by Gene Clusters

- **Gene clusters or operons encoding co-regulated genes are usually conserved, despite shuffling effects of evolution**

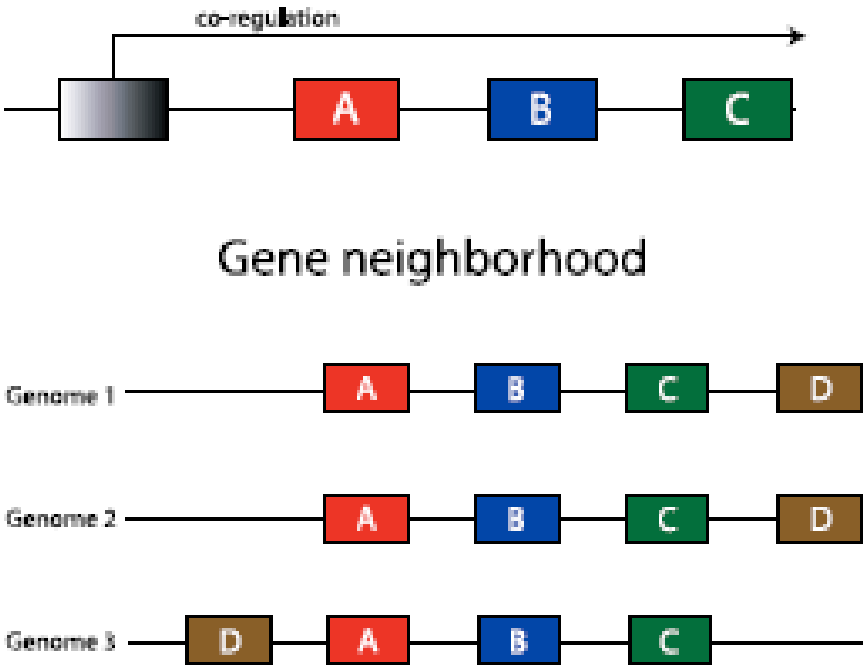


Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

⇒ Find conserved gene clusters

- Predict the genes to interact & form operons

PPI Prediction by Phylogenetic Profiling

- **Components of complexes and pathways should be present simultaneously in order to perform their functions**
- Functionally linked and interacting proteins co-evolve and have orthologs in the same subset of fully sequenced organisms

Proteins	Genomes EC	HI	BS
P1	0	1	1
P2	0	0	1
P3	1	0	0
P4	0	1	1

→ P1 and P4 are functionally linked

Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

PPI Prediction by Rosetta Stone

- **Some interacting proteins have homologs in other genomes that are fused into one protein chain, a so-called Rosetta Stone protein**
- Gene fusion occurs to optimize co-expression of genes encoding for interacting proteins

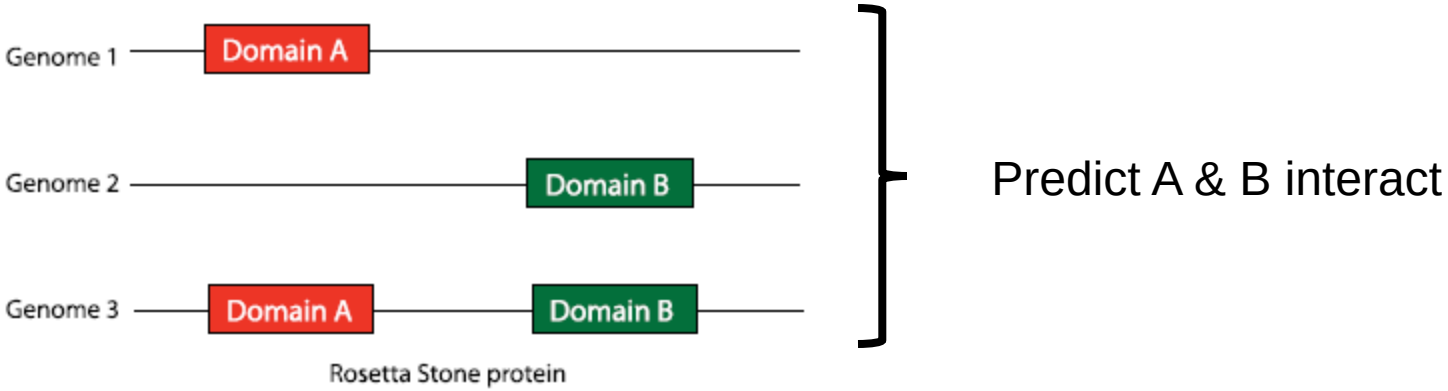


Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

PPI Prediction by Seq Co-Evolution

- **Interacting proteins co-evolve**
  - Changes in one protein leading to loss of function are compensated by correlated changes in another protein

- Co-evolution is quantified by correlation of distance matrices used to construct the trees

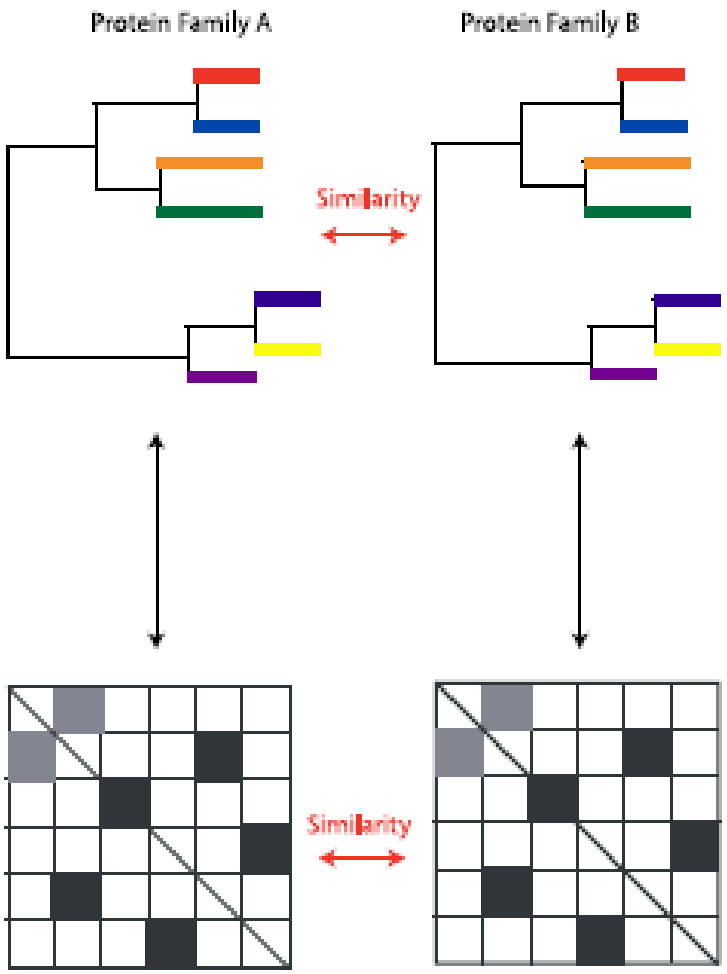
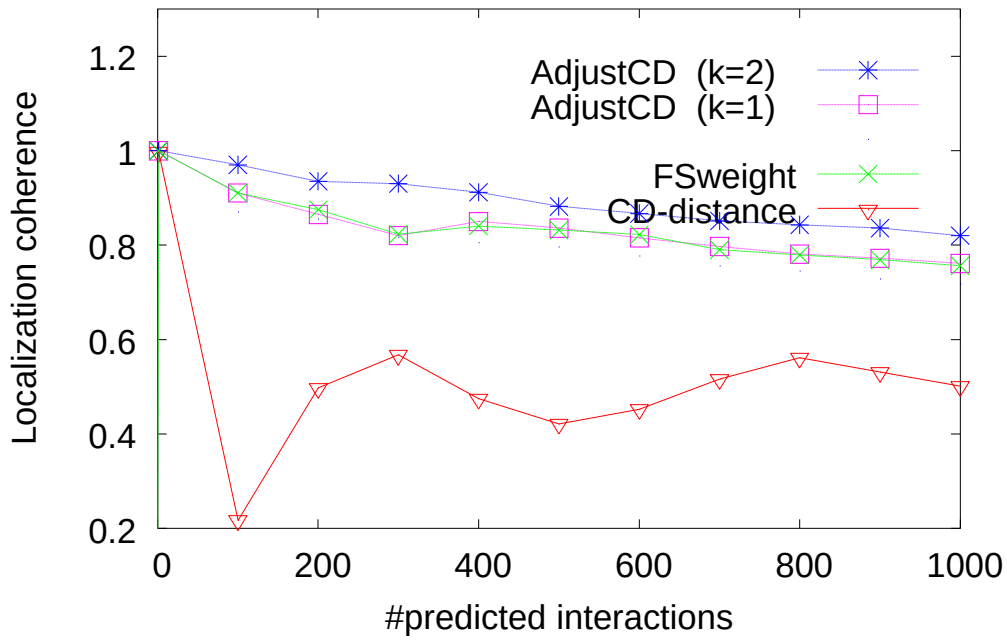


Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

Liu et al. *GIW2008*, pp. 138-149

PPI Prediction by Iterated CD-Distance

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



$$wL^k(u,v) = \frac{\sum_{x \in Nu \cap Nv} w_L^{k-1}(u,x) + \sum_{x \in Nu \cap Nv} w_L^{k-1}(v,x)}{\sum_{x \in Nu} w_L^{k-1}(u,x) + \lambda^k{}_u + \sum_{x \in Nv} w_L^{k-1}(v,x) + \lambda^k{}_v}$$

- Predict (u,v) interact if $wL^k(u,v)$ is large

References

- D Soh et al. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases**. *BMC Bioinformatics*, 11:449, 2010
- Chua & Wong. **Increasing the Reliability of Protein Interactomes.** *Drug Discovery Today*, 13(15/16):652--658, 2008
- Liu et al. **Assessing and predicting protein interactions using both local and global network topological metrics,** *GIW2008,* 138-149
- Shoemaker & Panchenko. **Deciphering protein-protein Interactions. Part II. Computational methods to predict protein and domain interaction partners***. PLoS Computational Biology*, 3(4):e43, 2007

NUS National University of Singapore

Acknowledgements

Kenny Chua

Difeng Dong

Wilson Goh

Donny Soh

- Singapore MOE scholarship
- A*STAR AGS & AIP scholarships
- A*STAR SERC PSF grant
- NRF CRP grant
- Wellcome Trust scholarship

Agency for Science, Technology and Research SINGAPORE

NATIONAL RESEARCH FOUNDATION
Prime Minister's Office, Republic of Singapore

wellcome trust