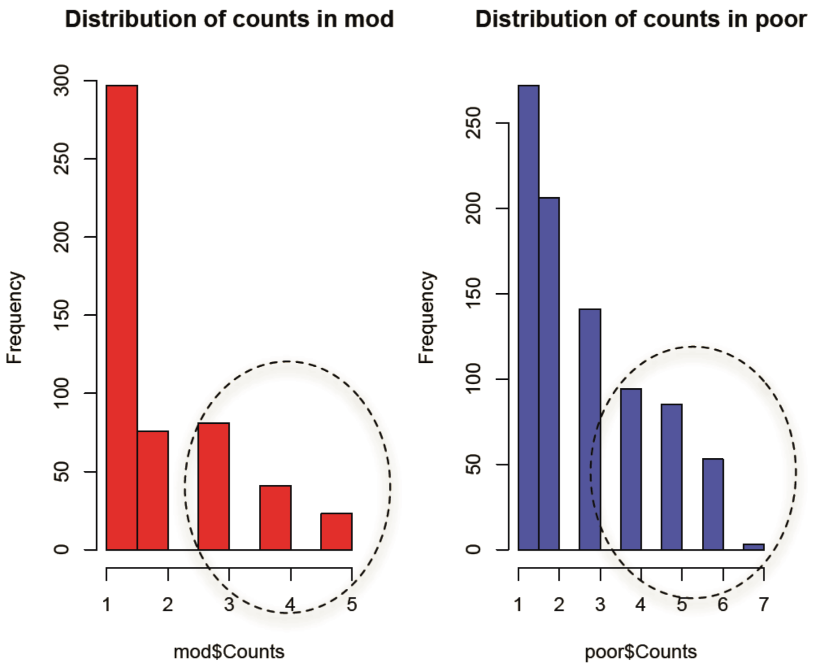
CS4220: Knowledge Discovery Methods for Bioinformatics

Unit 4: Proteomic Profile Analysis

Wong Limsoon

NUS
National University
of Singapore

Delivering more powerful proteomic profile analysis



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

Basic Proteomic Profile Analysis

NUS
National University
of Singapore

Typical Proteomic MS Experiment

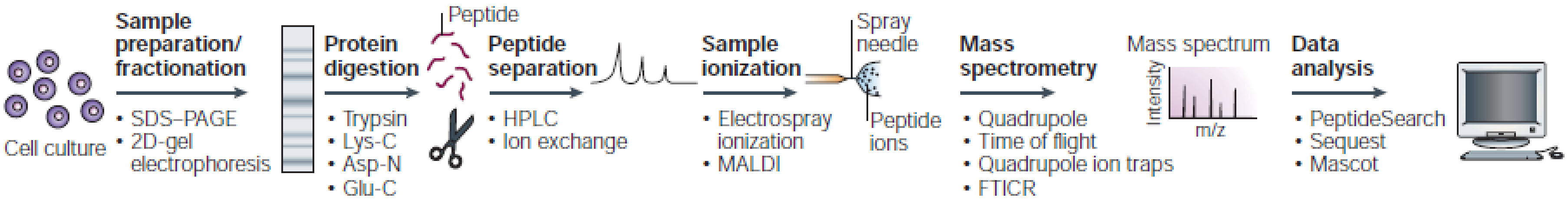


Figure 1 | **The mass-spectrometry/proteomic experiment.** A protein population is prepared from a biological source — for example, a cell culture — and the last step in protein purification is often SDS–PAGE. The gel lane that is obtained is cut into several slices, which are then in-gel digested. Numerous 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.

See also http://www.slideshare.net/joachimjacob/bits-introduction-to-mass-spec-data-generation

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

Diagnosis Using Proteomics

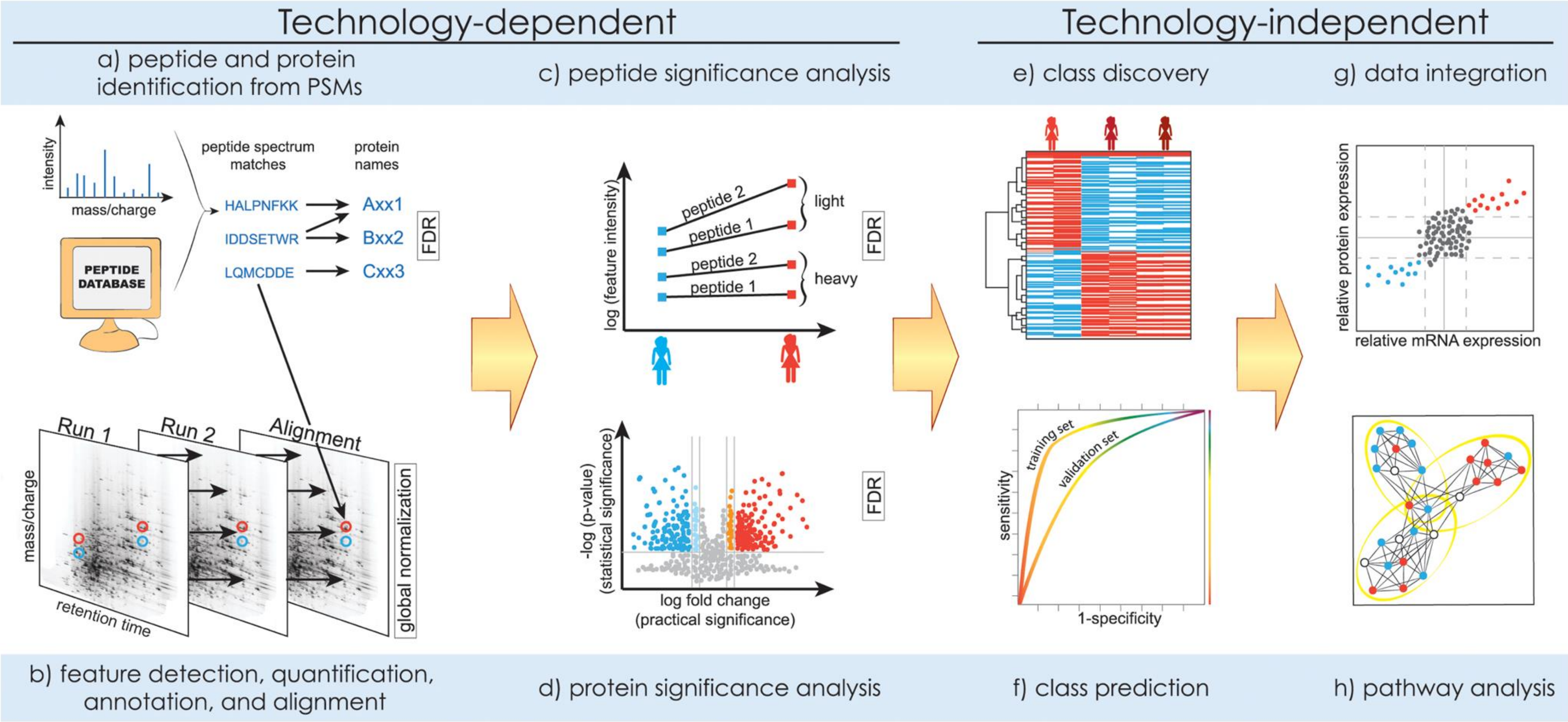


Image credit: Kall and Vitek, *PLoS Comput Biol* , 7(12): e1002277, 2011

A rather nice set of proteomic profiles of leukemia patients

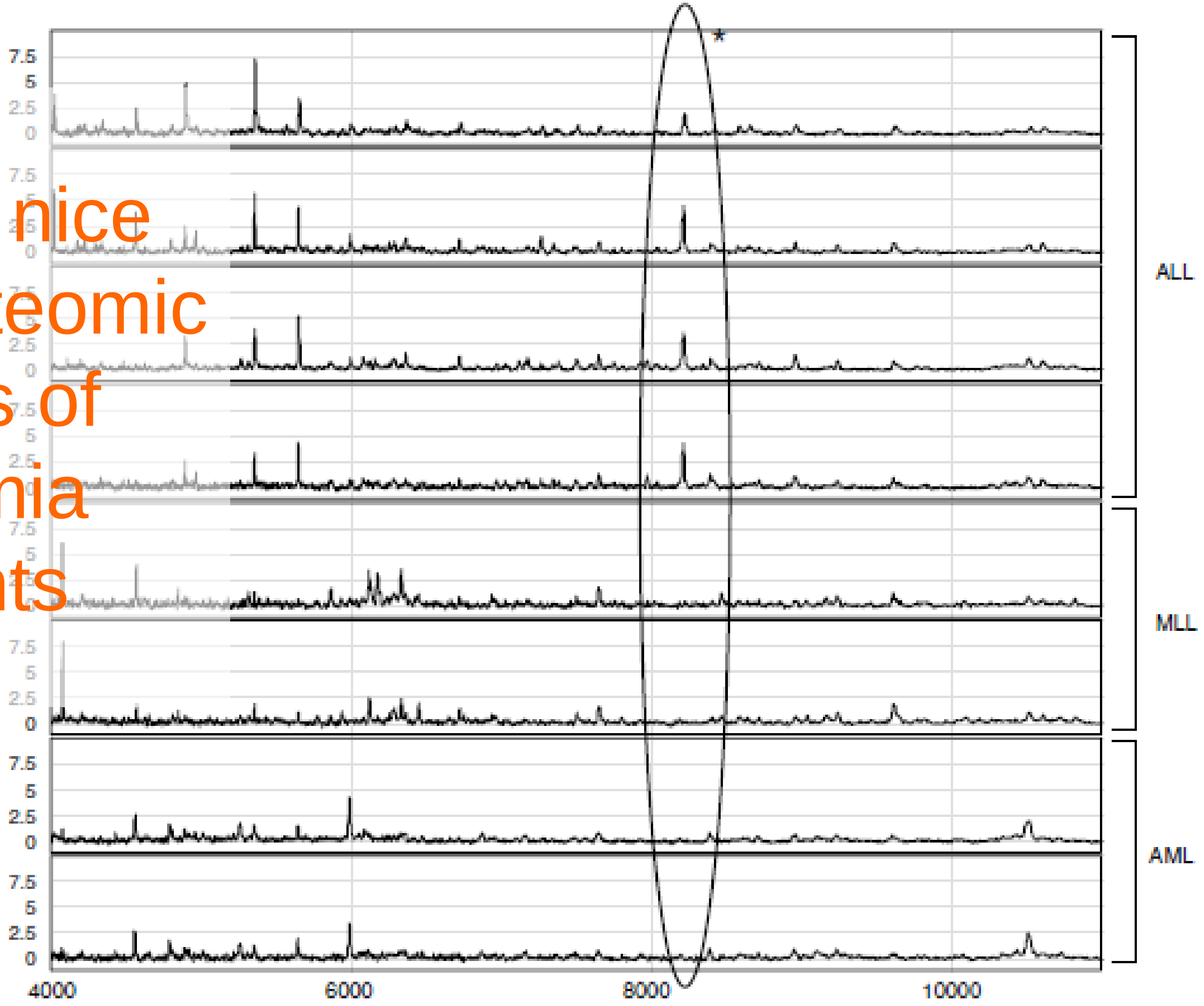
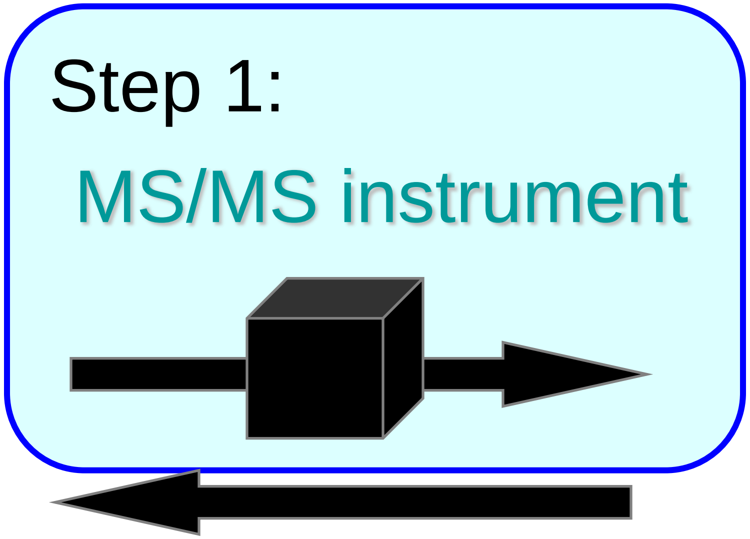


Figure 1 Spectra from SELDI-TOF MS analysis of REH, 697, MV4;11, and Kasumi cell lines. Protein (4 μg) from each cell type was analyzed on SAX2 ProteinChip® Arrays. ALL cell lines shown are REH and 697, the MLL cell line is MV4;11, and the AML cell line is Kasumi. The asterisk indicates the differentially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

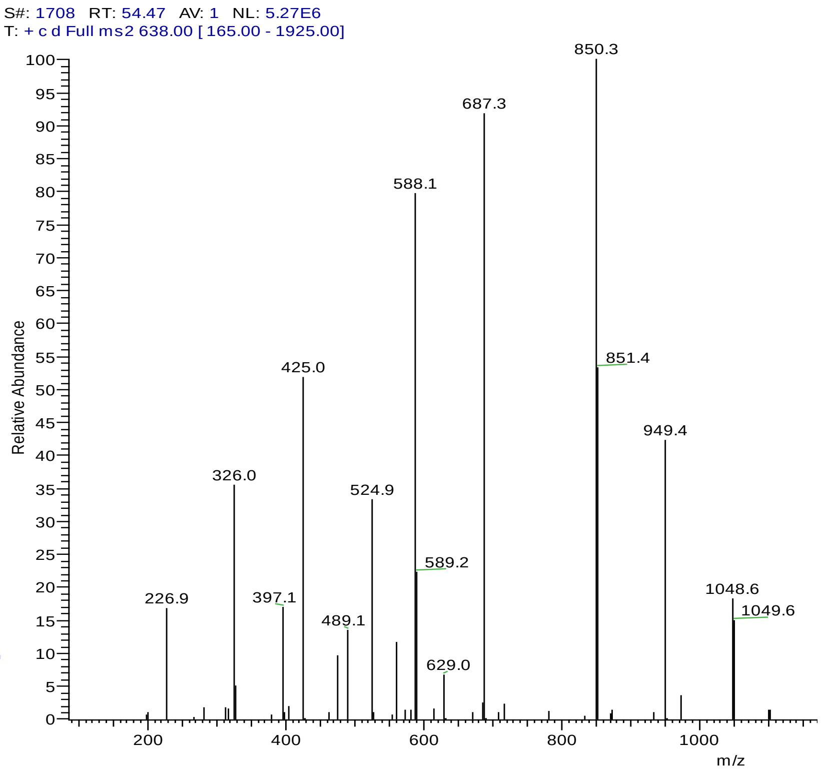
MS/MS instrument

Database search

- Sequest, Mascot, InSpect

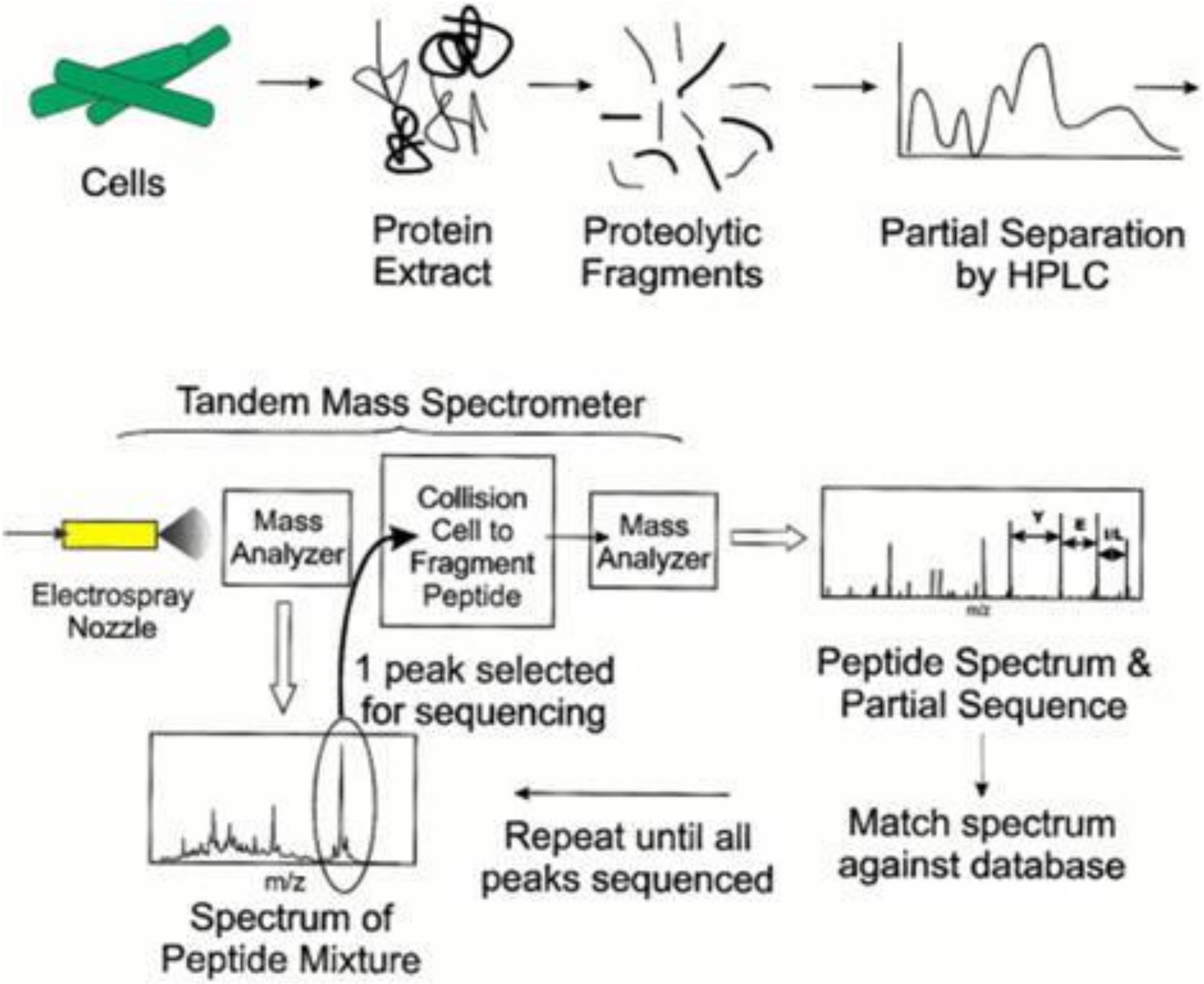
de Novo interpretation

- Lutefisk, Peaks, PepNovo



Source: Leong Hon Wai

Tandem Mass-Spectrometry


Cells
Protein Extract
Proteolytic Fragments
Partial Separation by HPLC
Tandem Mass Spectrometer
Electrospray Nozzle
Mass Analyzer
Collision Cell to Fragment Peptide
Mass Analyzer
1 peak selected for sequencing
m/z
Spectrum of Peptide Mixture
Repeat until all peaks sequenced
m/z
Peptide Spectrum & Partial Sequence
Match spectrum against database

Source: Leong Hon Wai

Breaking Protein into Peptides, and Peptides into Fragment Ions

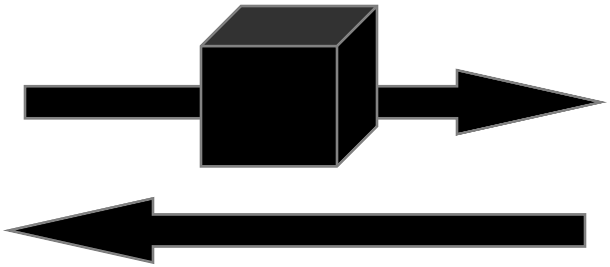
- **Proteases, e.g. trypsin, break protein into peptides**
- **A Tandem Mass Spectrometer further breaks the peptides down into fragment ions and measures the mass of each piece**
- **Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones**
- **Mass Spectrometer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

Source: Leong Hon Wai

Peptide Identification by Mass Spec

S
e
q
u
e
n
c
e

MS/MS instrument



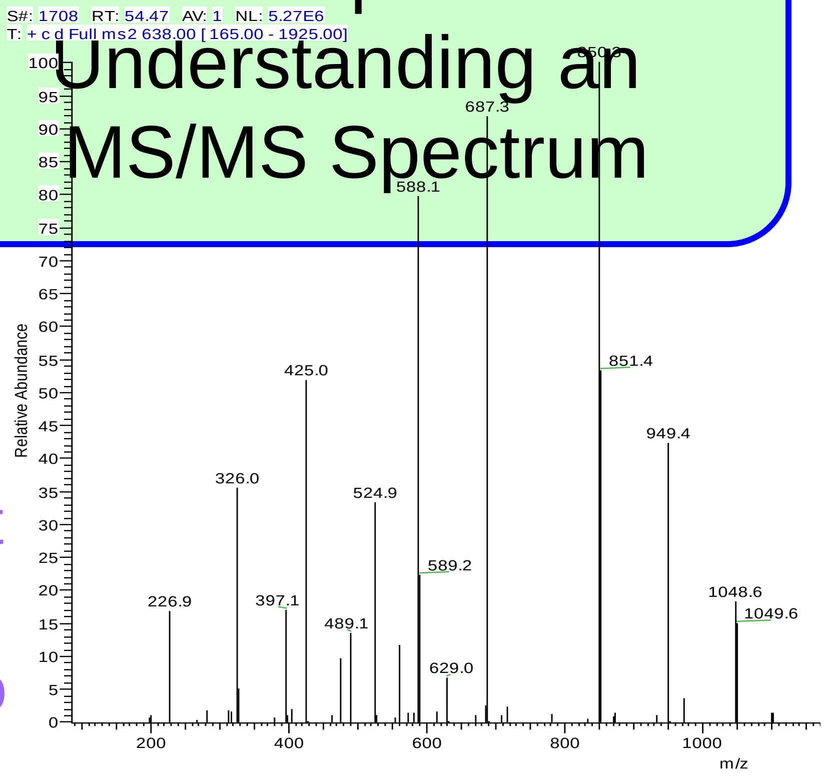
Database search

- Sequest, Mascot, InSpect

de Novo interpretation

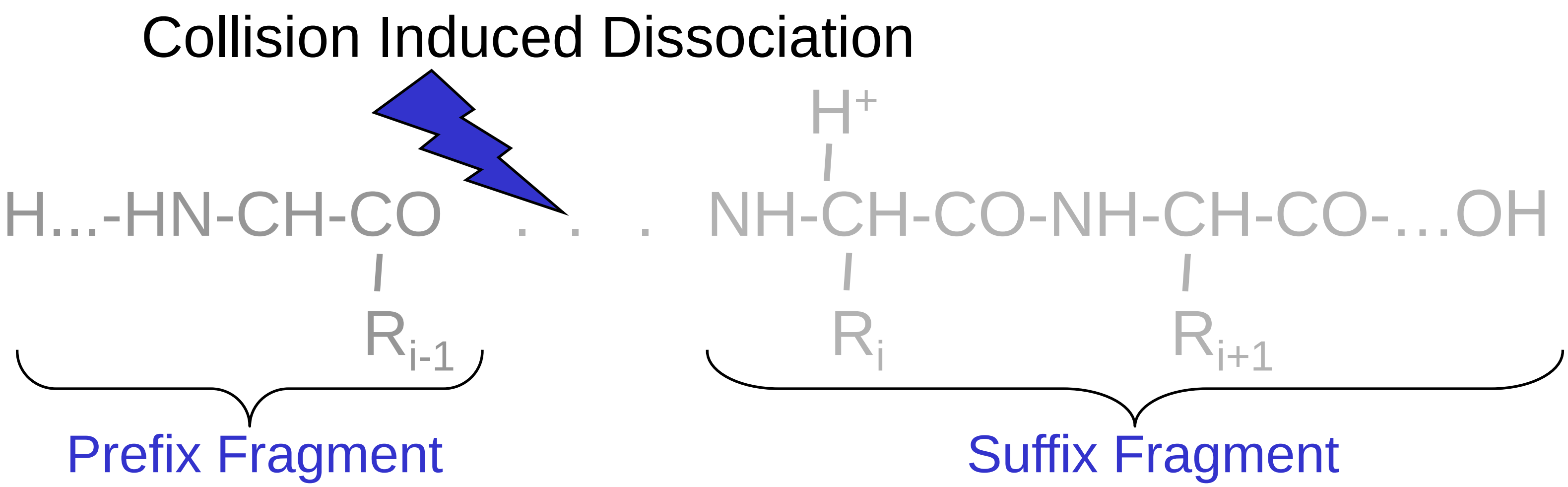
- Lutefisk, Peaks, PepNovo

Step 2:
Understanding an MS/MS Spectrum



Source: Leong Hon Wai

Peptide Fragmentation



Collision Induced Dissociation

- **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

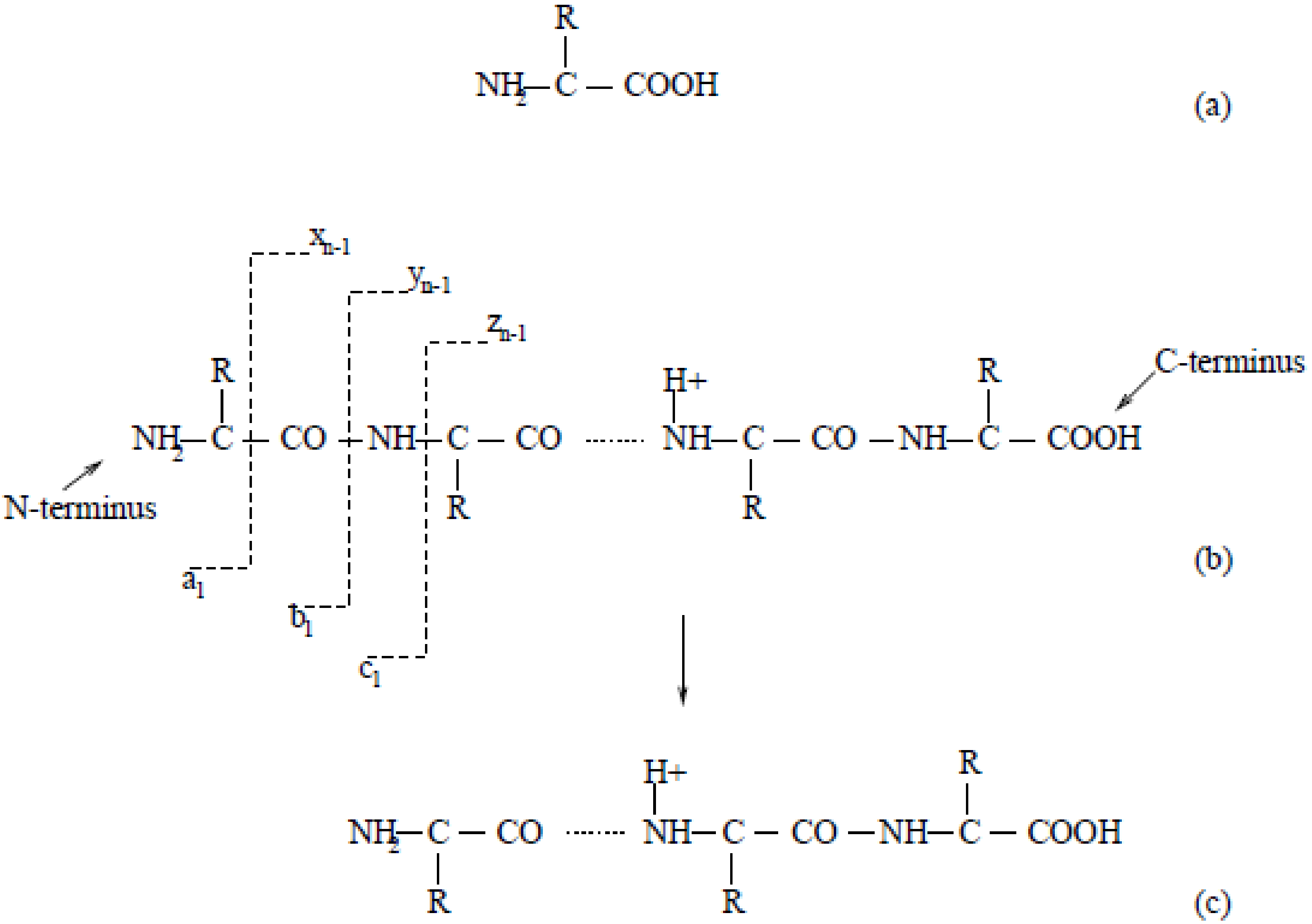
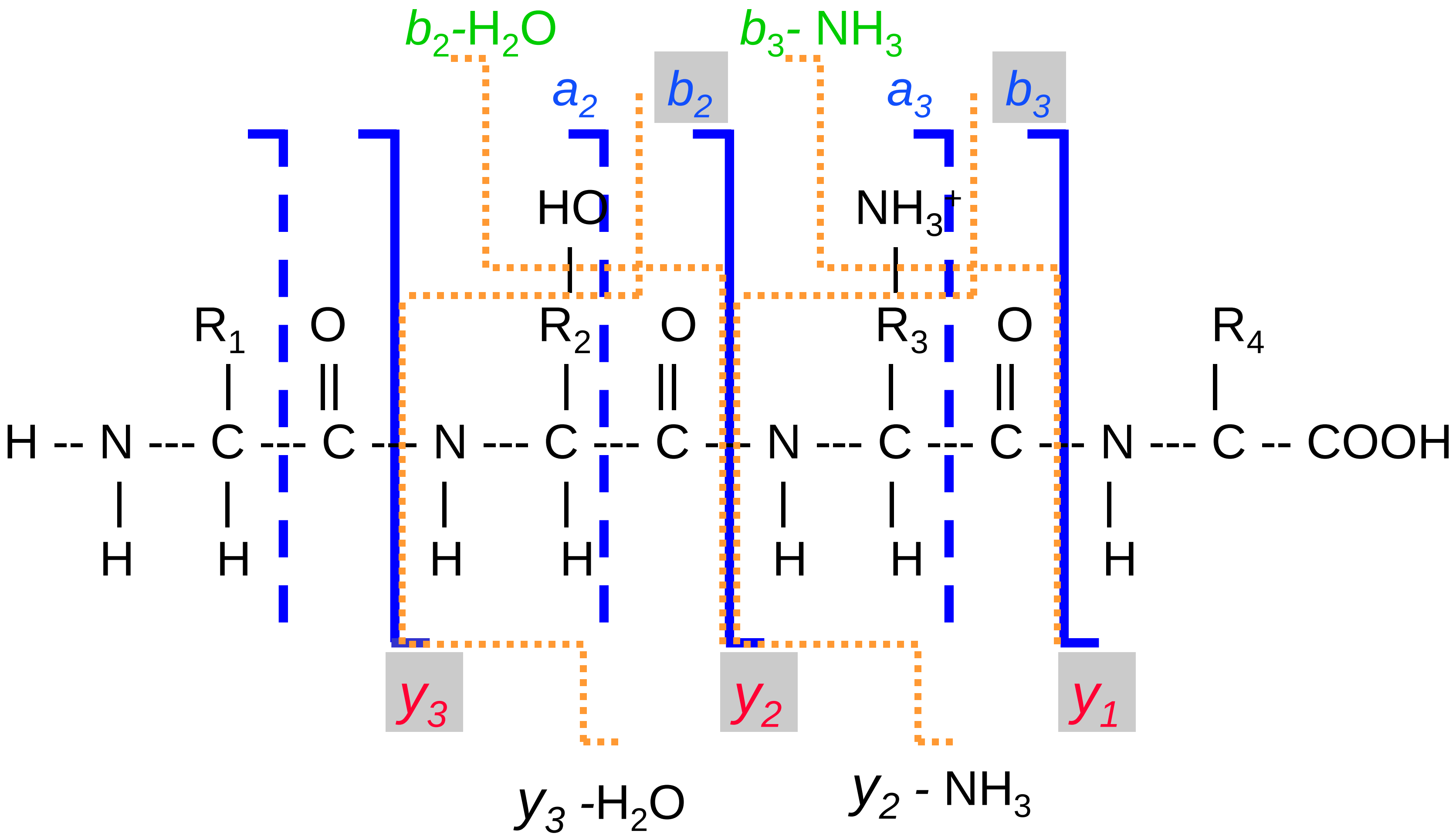


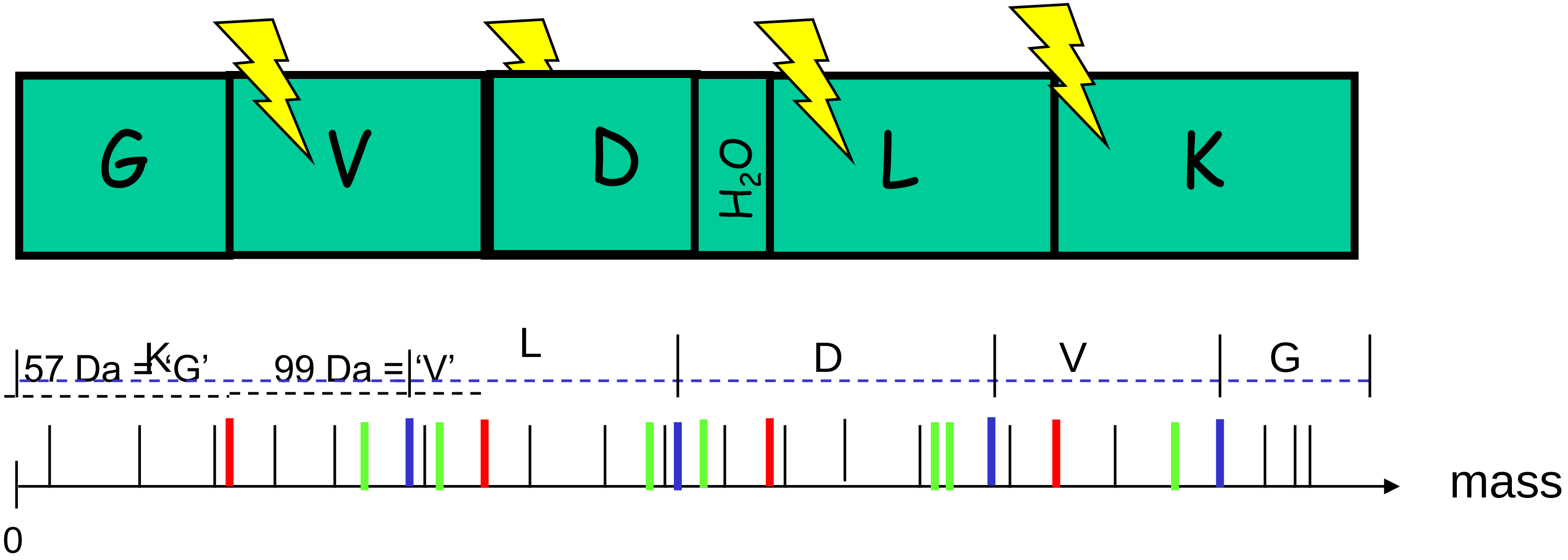
Figure 1: (a) The structure of an amino-acid. (b) An ionized peptide. (c) y^{+}_{n-1} ion

… 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

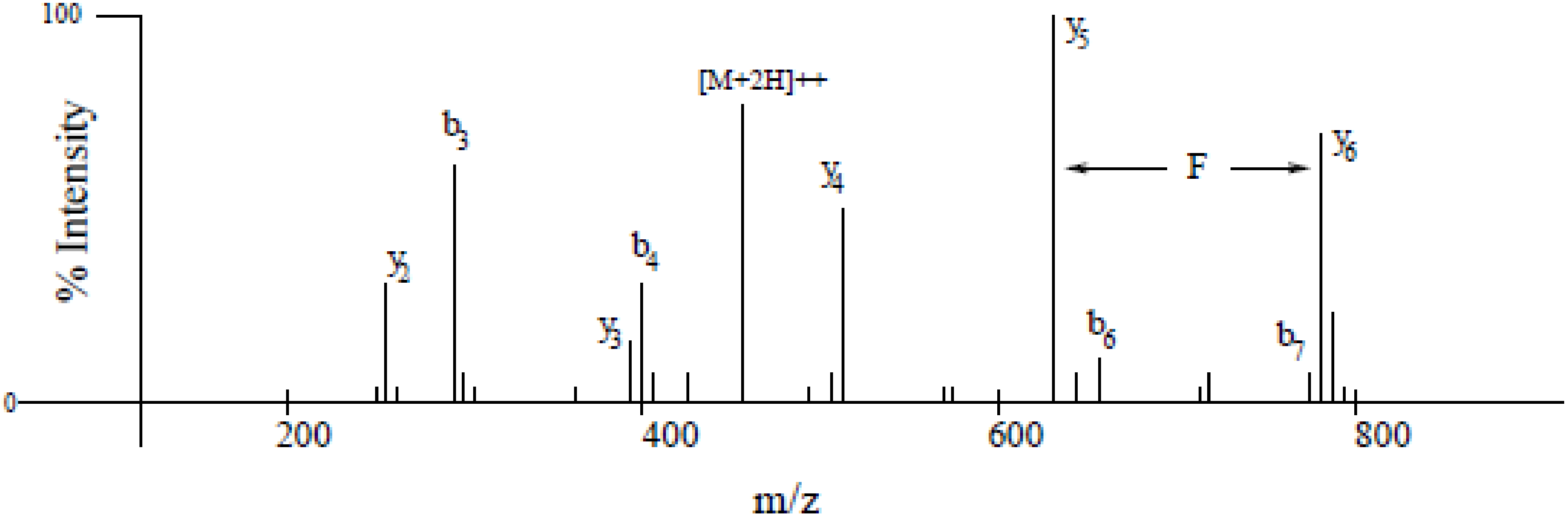
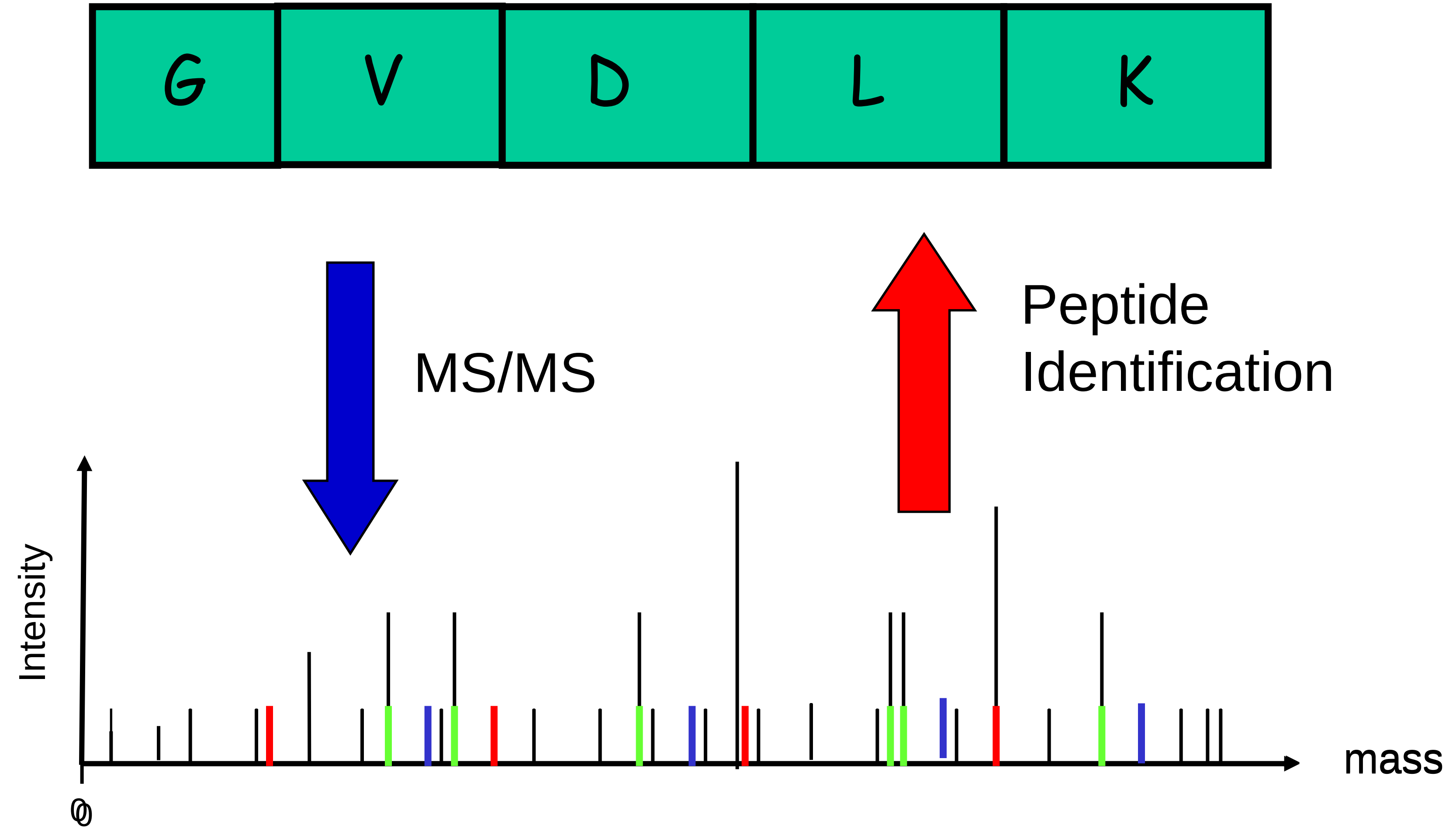


Figure 2: MS/MS spectrum for peptide SGFLEEDK.

Protein Identification with MS/MS

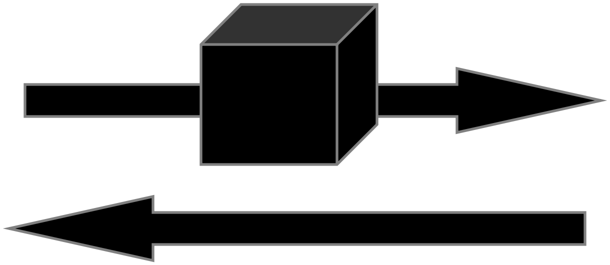


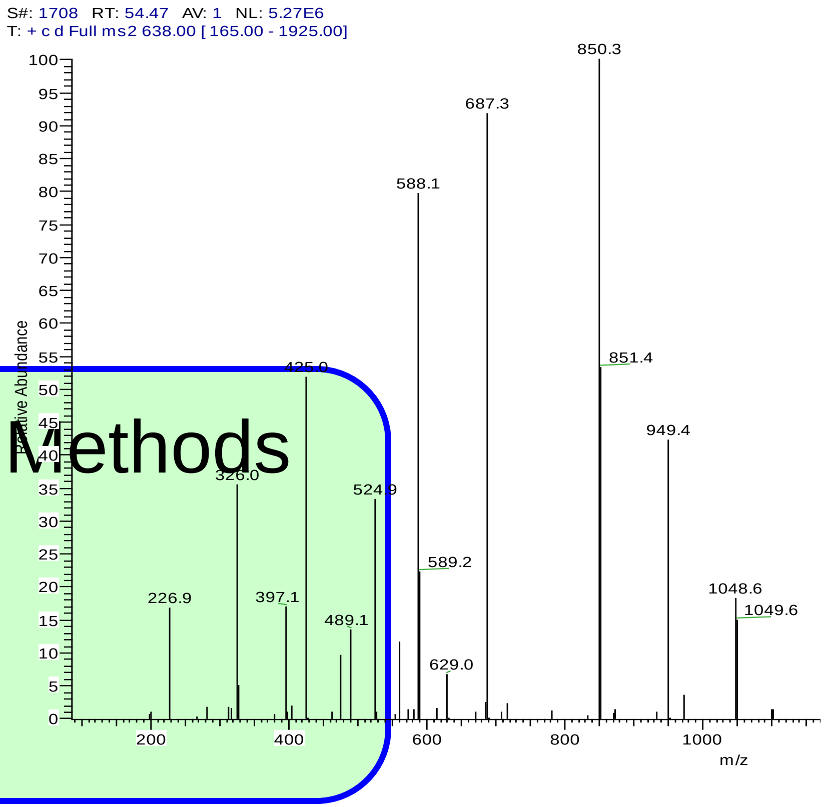
Source: Leong Hon Wai

Peptide Identification by Mass

Sequence

MS/MS instrument





Step 3: Computational Methods

Database search

- Sequest, Mascot

de Novo interpretation

- Lutefisk, Peaks, PepNovo

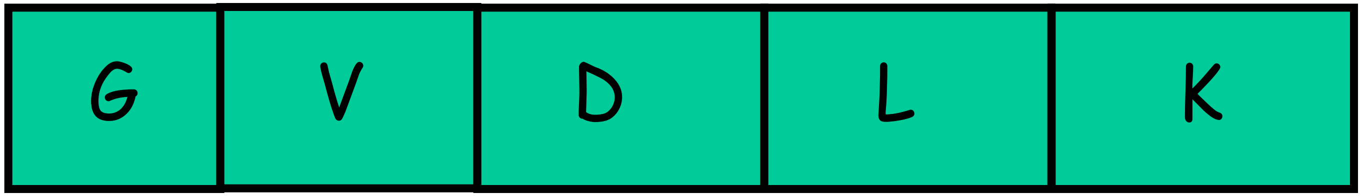
Source: Leong Hon Wai

Database Search Algorithms

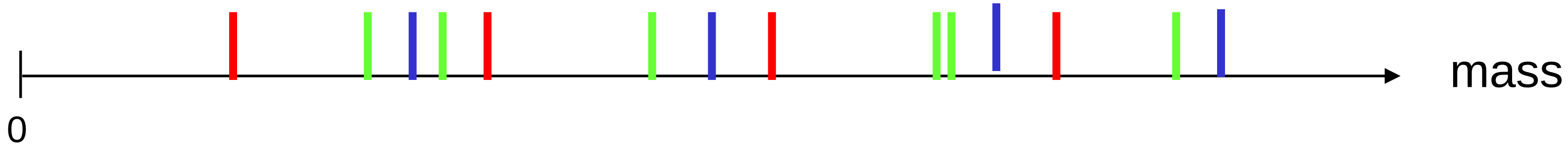
- **Database search**
  - Used for spectrum from known peptides
  - Rely on completeness of database
- **General Approach**
  - Match given spectrum with known peptide
  - Enhanced with advanced statistical analysis and complex scoring functions
- **Methods**
  - SEQUEST, MASCOT, InsPecT, Paragon

Theoretical Spectrum for a Peptide

- **Given this peptide**



- **Its theoretical spectrum is**

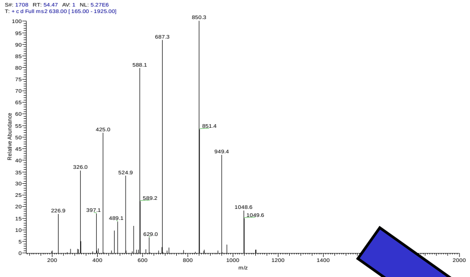


- **Theoretical spectrum is dependent on**
  - Set of ion-types considered
  - Larger if multi-charge ions are considered

Source: Leong Hon Wai

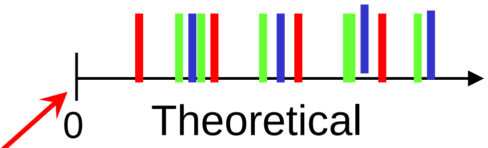
Database Search Algorithm

Database Search



Database of known peptides

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



Theoretical spectrum

Match

Matching Score for this peptide

Repeat for all the peptides in the Database

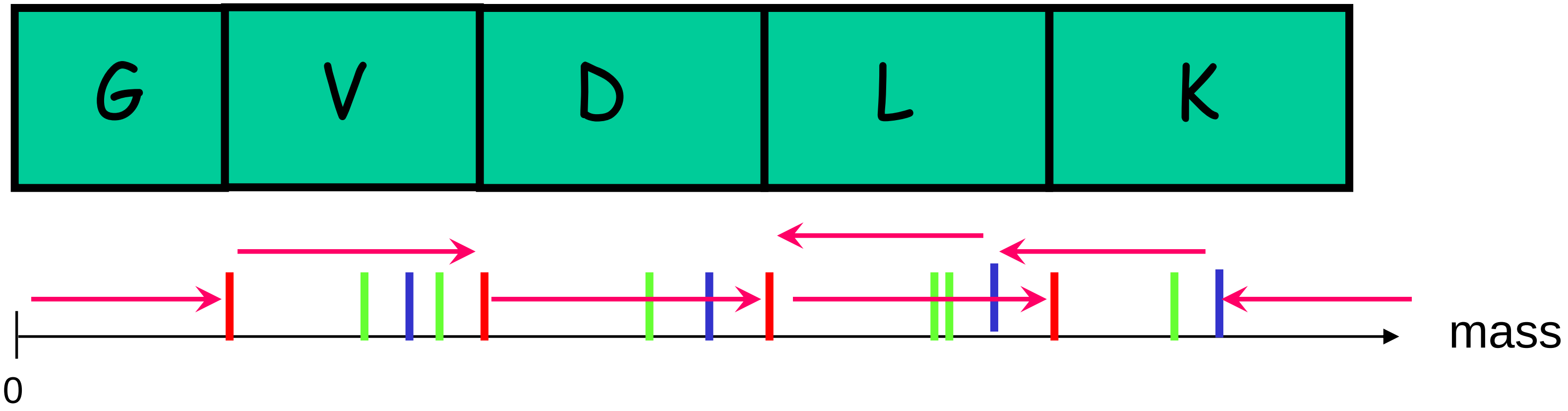
Source: Leong Hon Wai

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

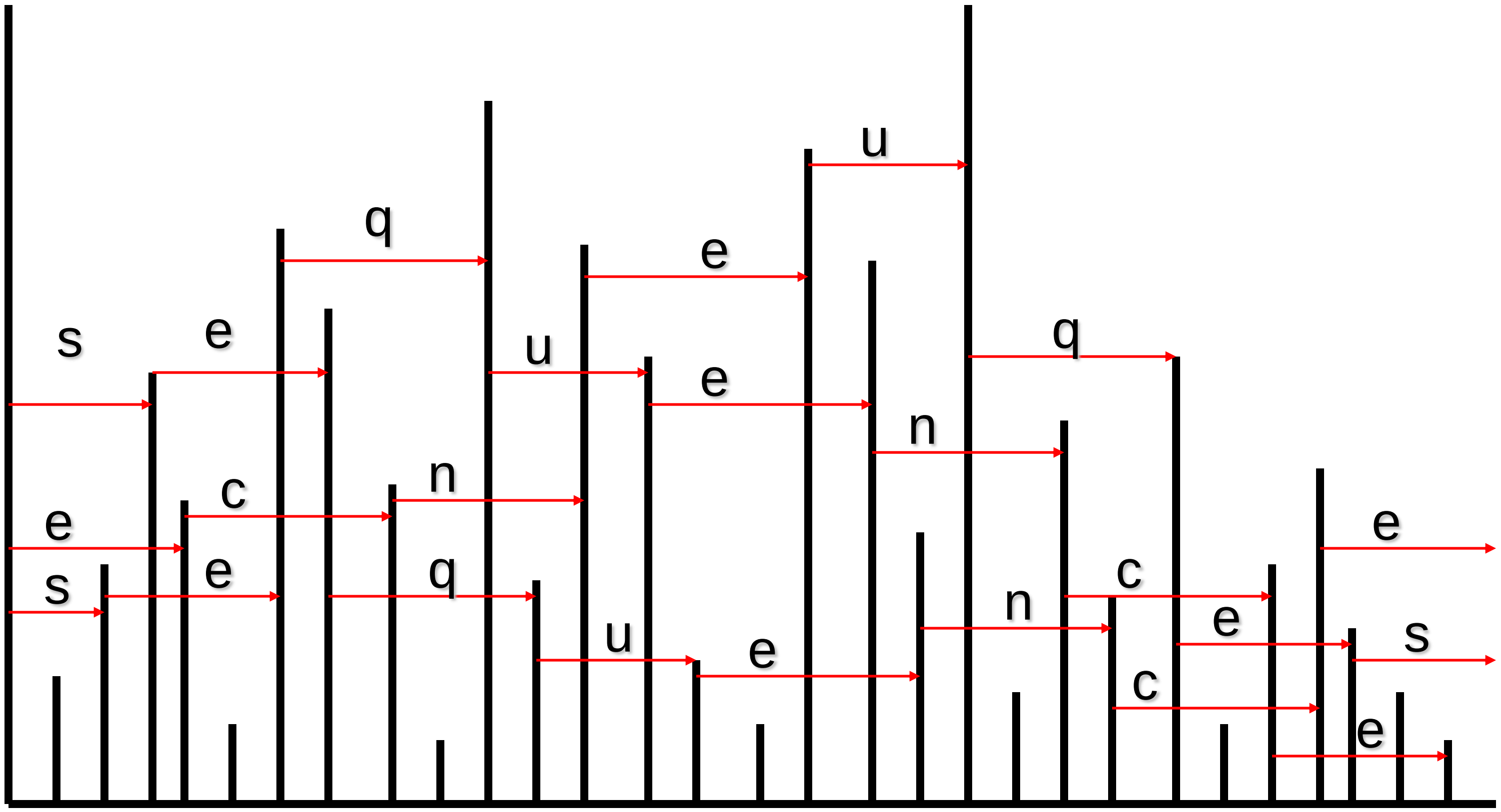
Spectrum Graph for a Peptide



- **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

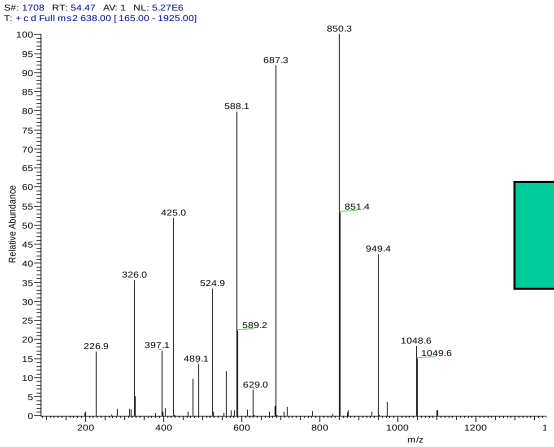
Building a Graph from a Spectrum

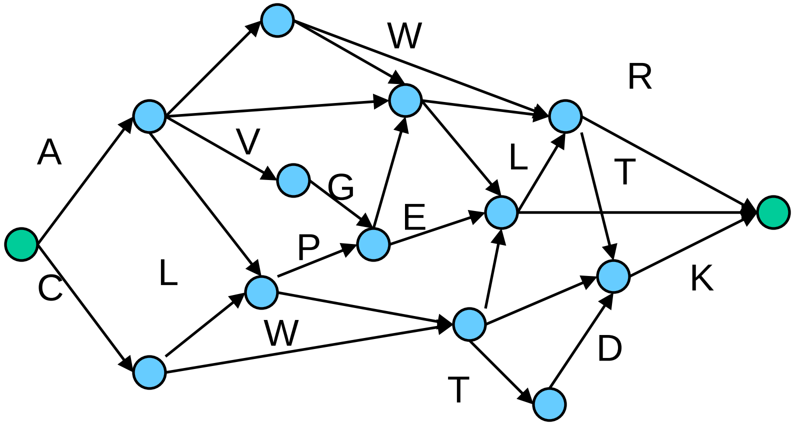


Source: Leong Hon Wai

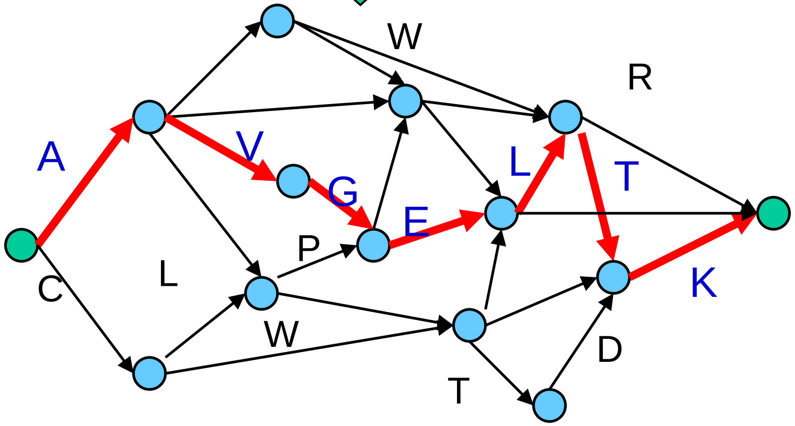
Frank, et al. “De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry”. J. Proteome Res. 6:114-123, 2007

De Novo Sequencing Algorithms



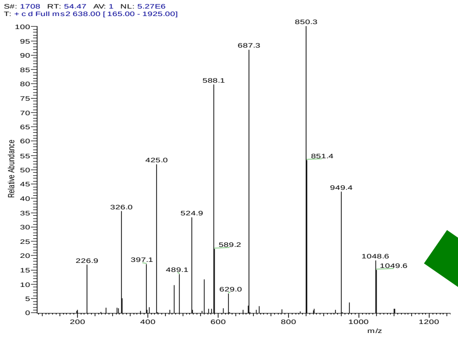


Find longest directed acyclic path



AVGELTK

De Novo vs. Database Search

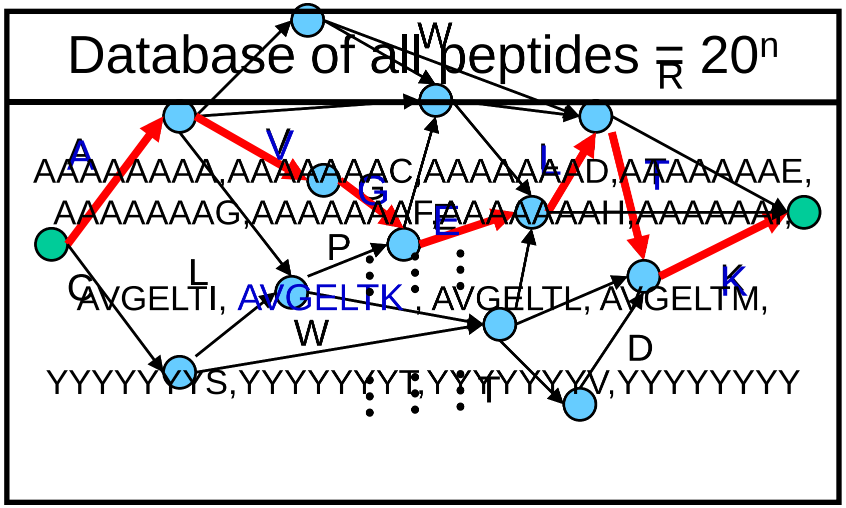


Database Search

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

AAAAAAAA,AAAAAAAC,AAAAAAAD,AAAAAAAE,
AAAAAAAG,AAAAAAAF,AAAAAAAH,AAAAAAAI,
AVGELTI, AVGELTK , AVGELTL, AVGELTM,
YYYYYYYS,YYYYYYYT,YYYYYYYV,YYYYYYYY

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

Common Issues in Proteomic Profile Analysis

NUS

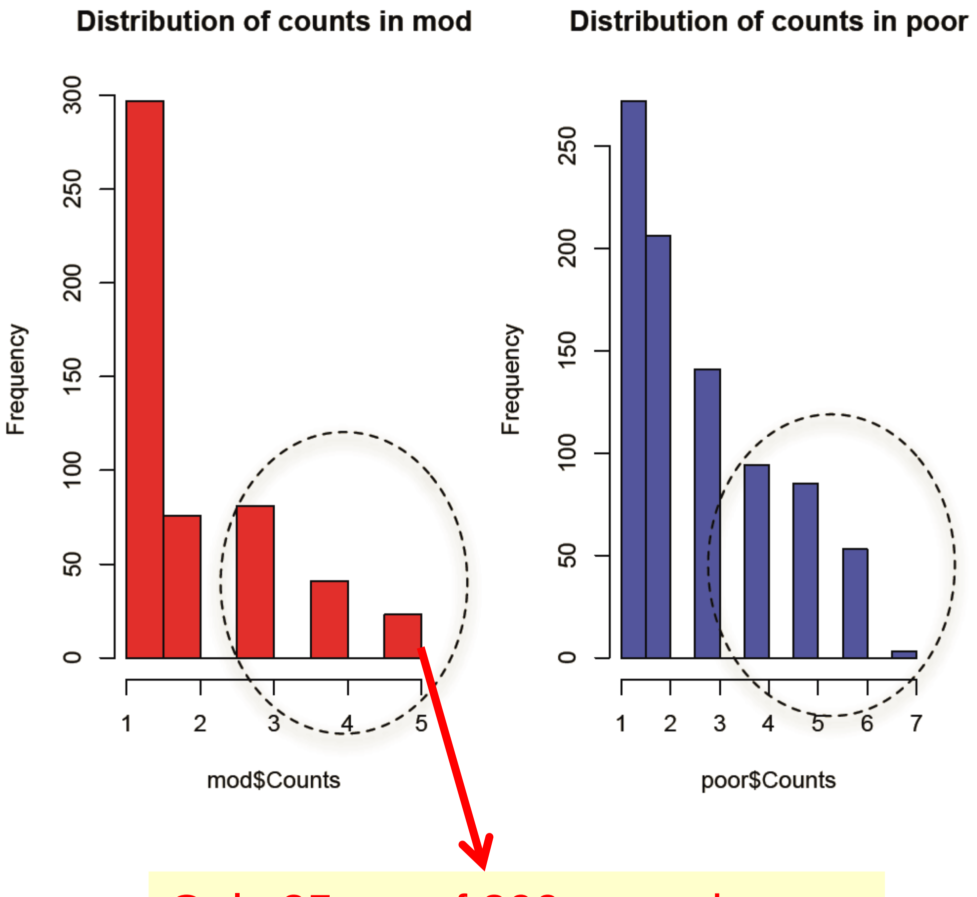
National University of Singapore

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**

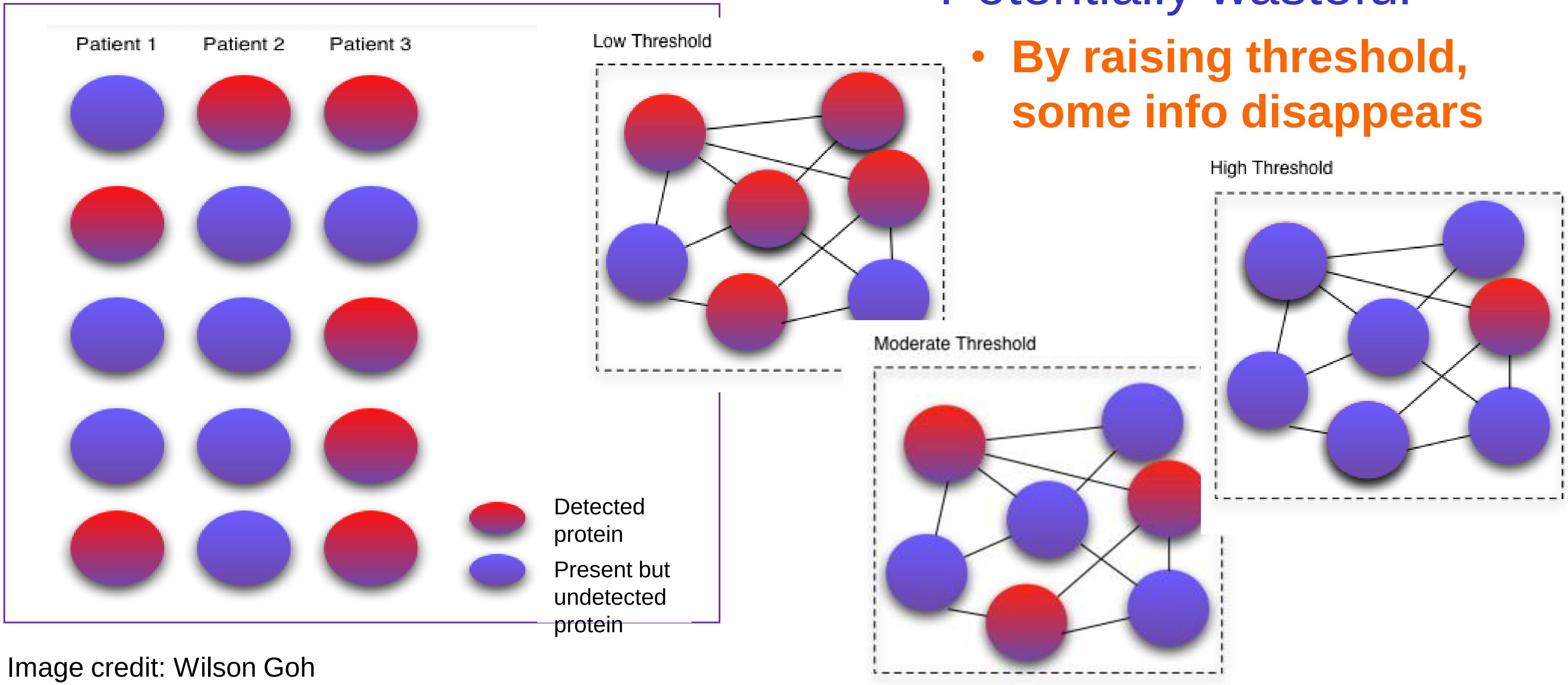
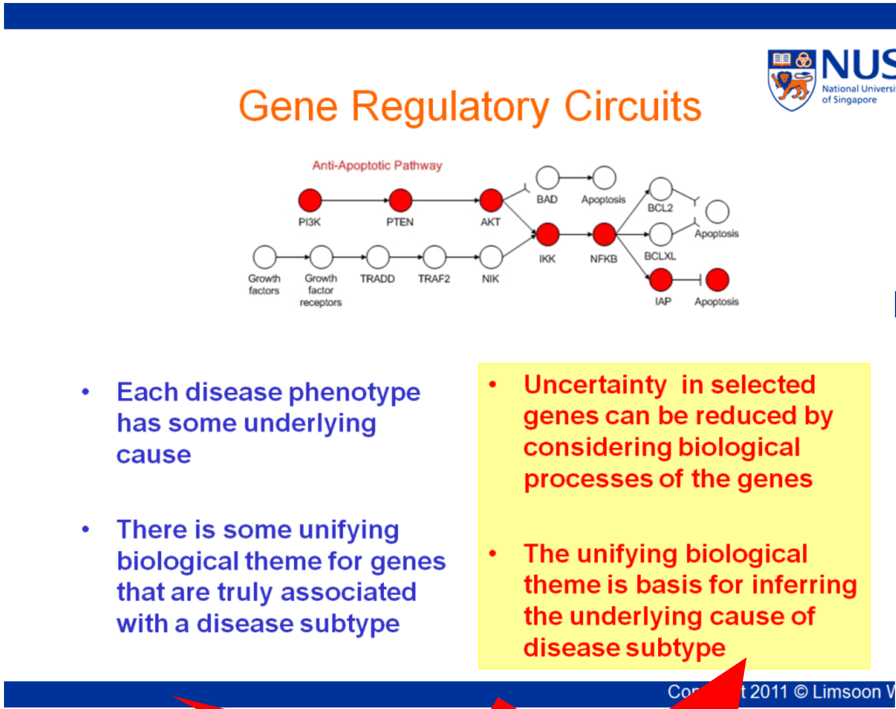


Image credit: Wilson Goh

Improving Consistency in Proteomic Profile Analysis

NUS
National University
of Singapore

An inspiration from gene expression profile analysis

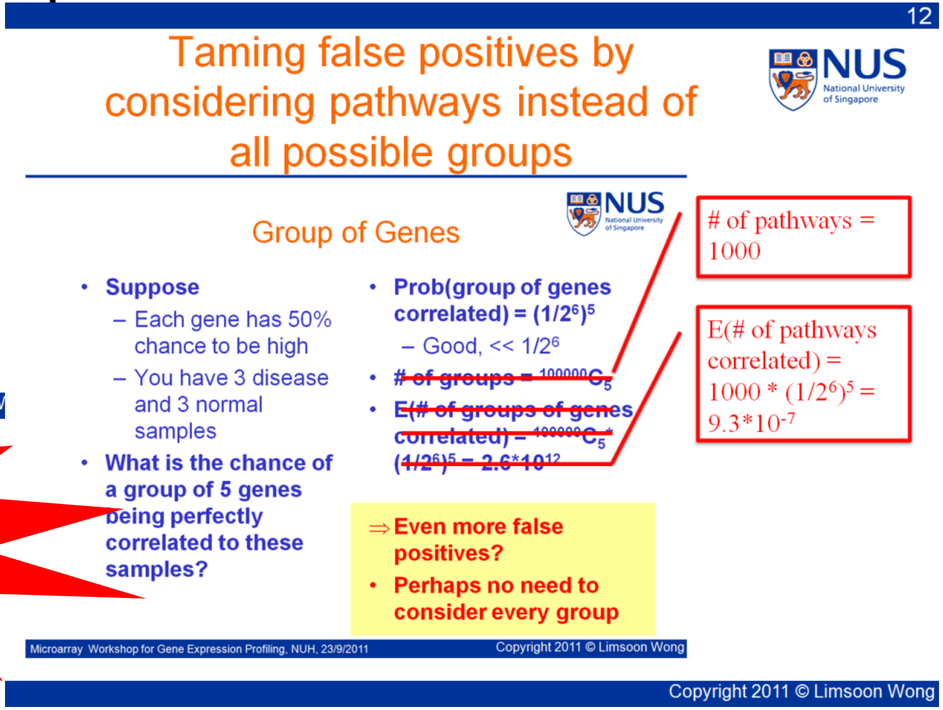


Gene Regulatory Circuits

- 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

Contextualization!



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) = ${}^{100000}C_5$ * $(1/2^6)^5$ = $2.6*10^{12}$**~~

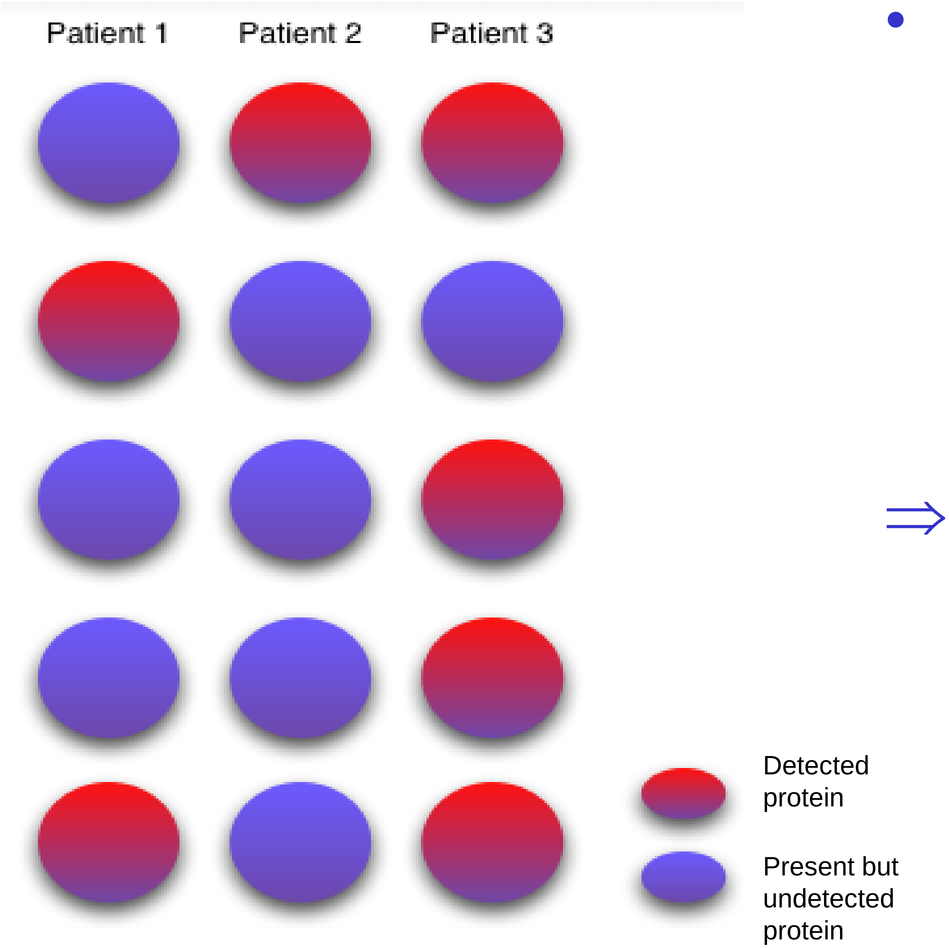
⇒ **Even more false positives?**

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

# of pathways = 1000

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

Intuitive Example



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

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

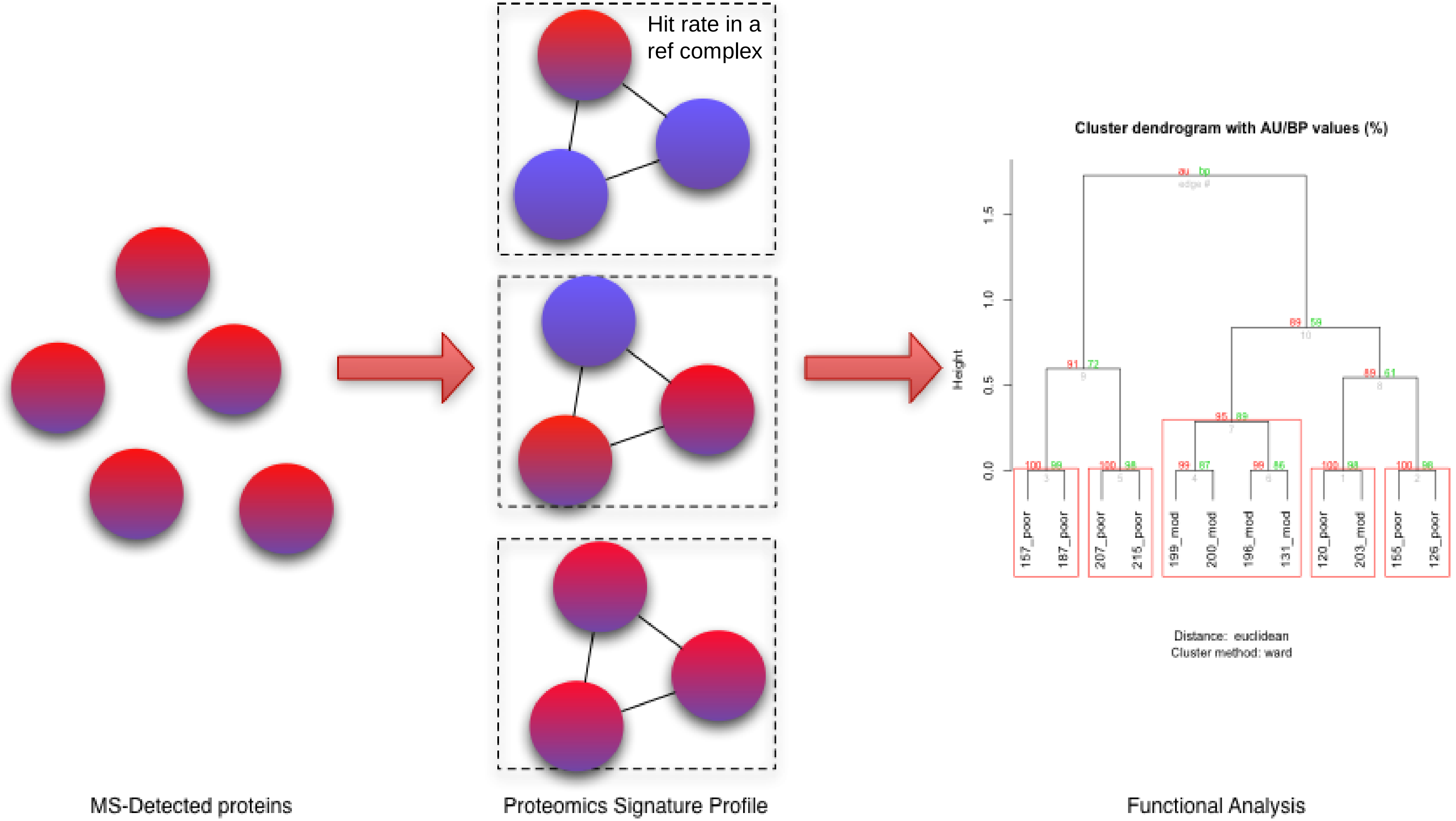
- Construct a profile based on complexes?

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*. 11(3):1571-1581, March 2012.

“Threshold-free” Principle of PSP



Hit rate in a ref complex

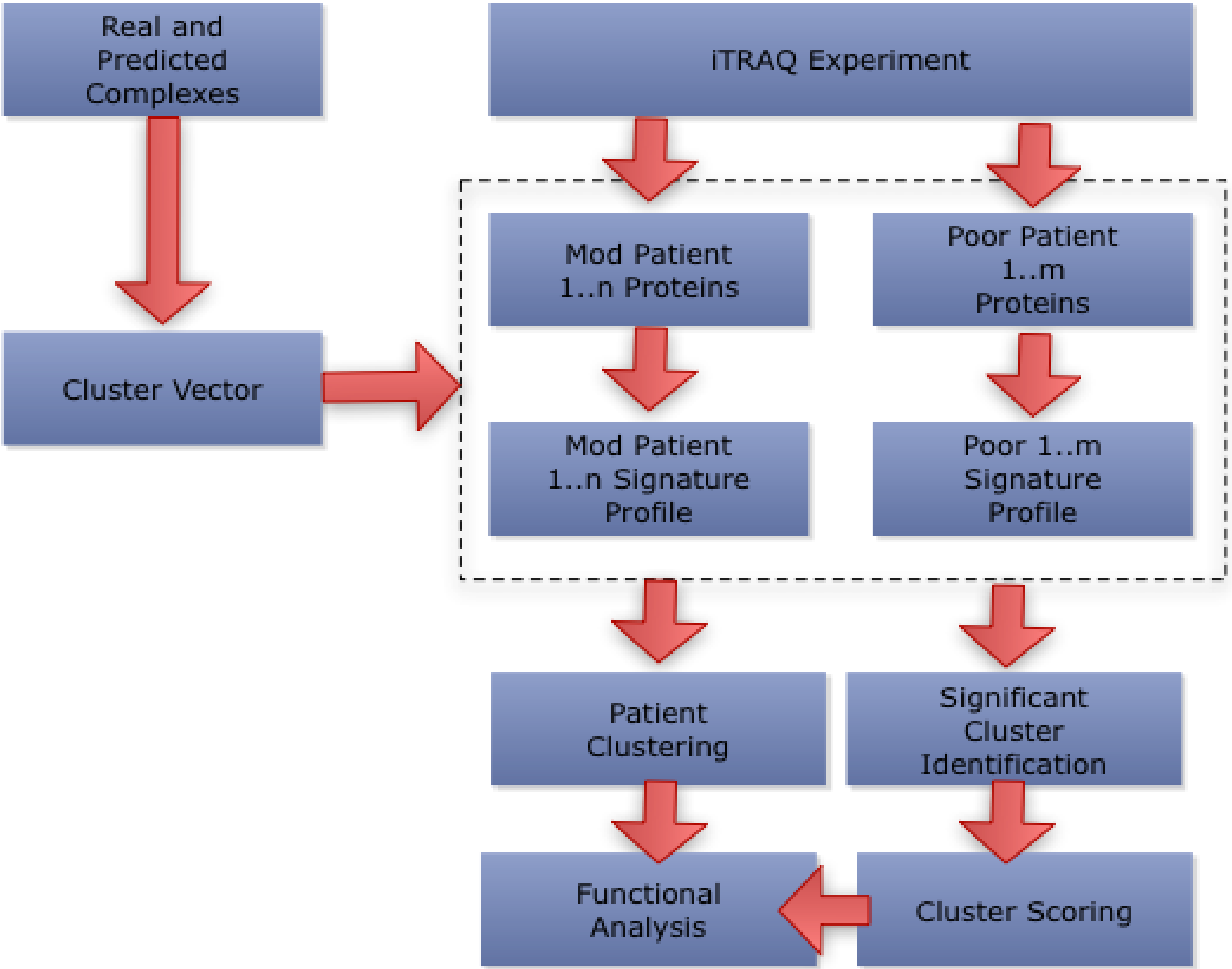
MS-Detected proteins

Proteomics Signature Profile

Functional Analysis

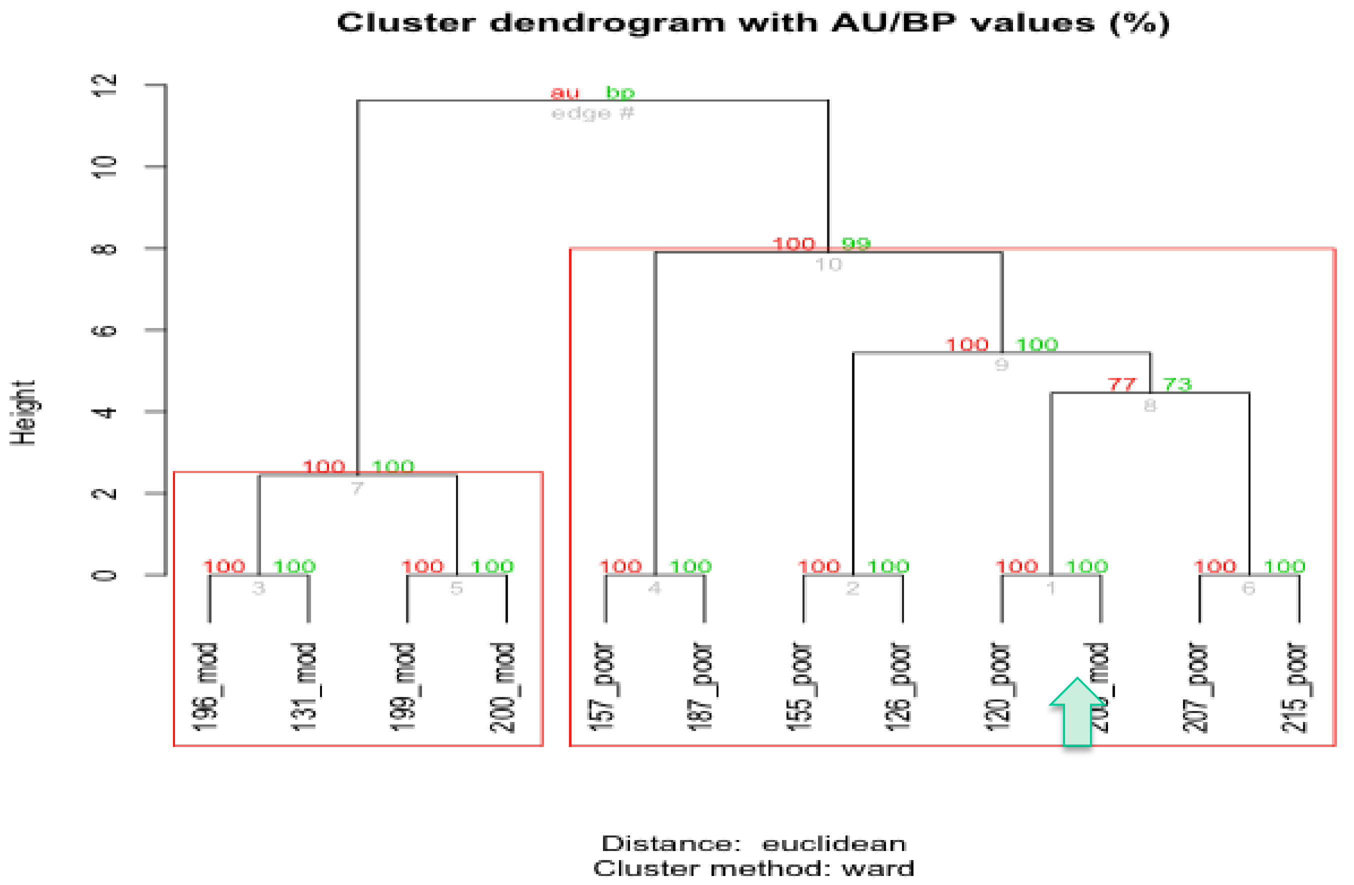
Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 11(3):1571-1581, March 2012.

Applying PSP to a HCC Dataset


Real and Predicted Complexes
iTRAQ Experiment
Mod Patient 1..n Proteins
Poor Patient 1..m Proteins
Cluster Vector
Mod Patient 1..n Signature Profile
Poor 1..m Signature Profile
Patient Clustering
Significant Cluster Identification
Functional Analysis
Cluster Scoring

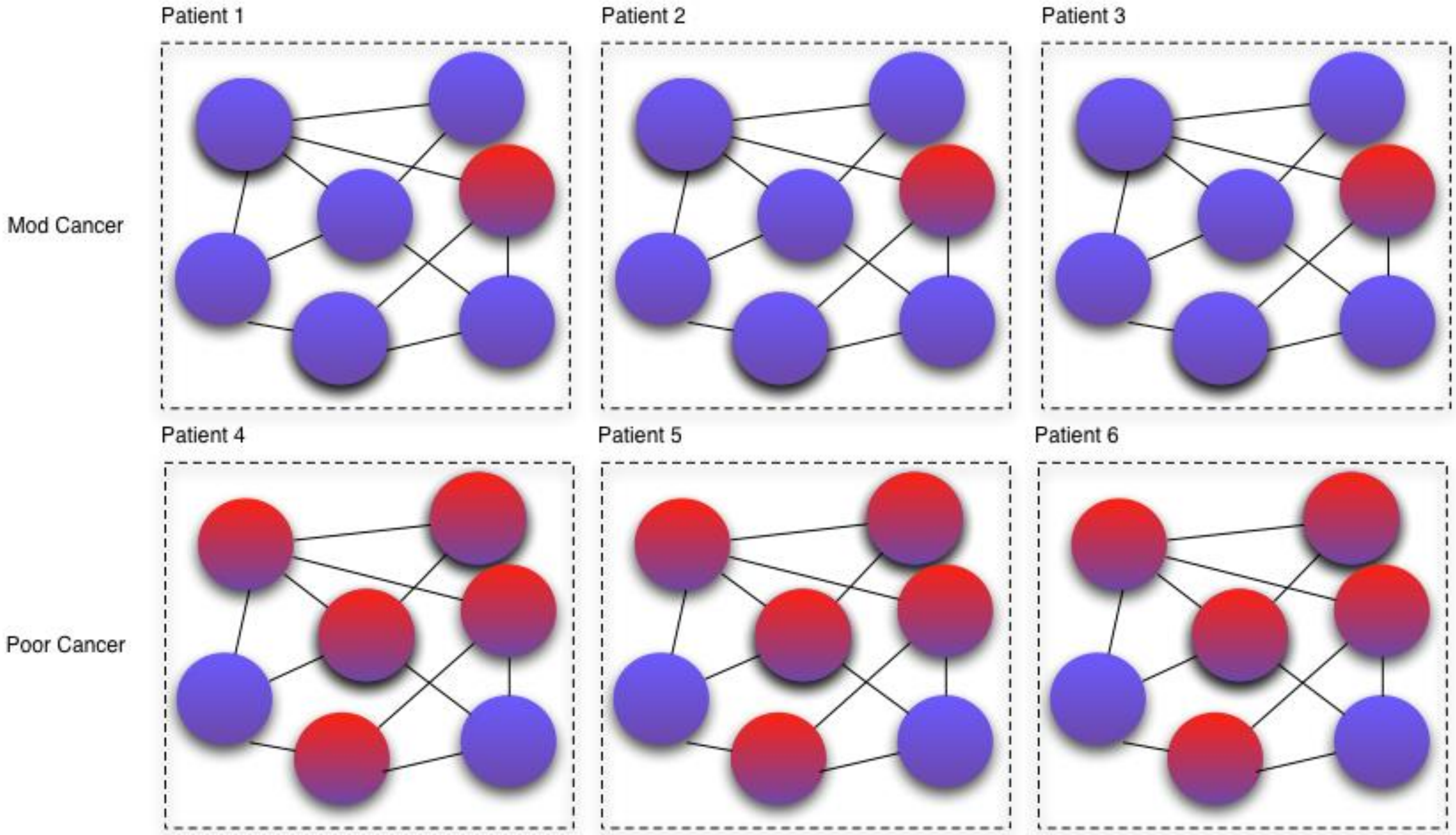
Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 11(3):1571-1581, March 2012

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*. 11(3):1571-1581, March 2012

Feature Selection



🔴 Detected Protein

🔵 Undetected Protein

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

where

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

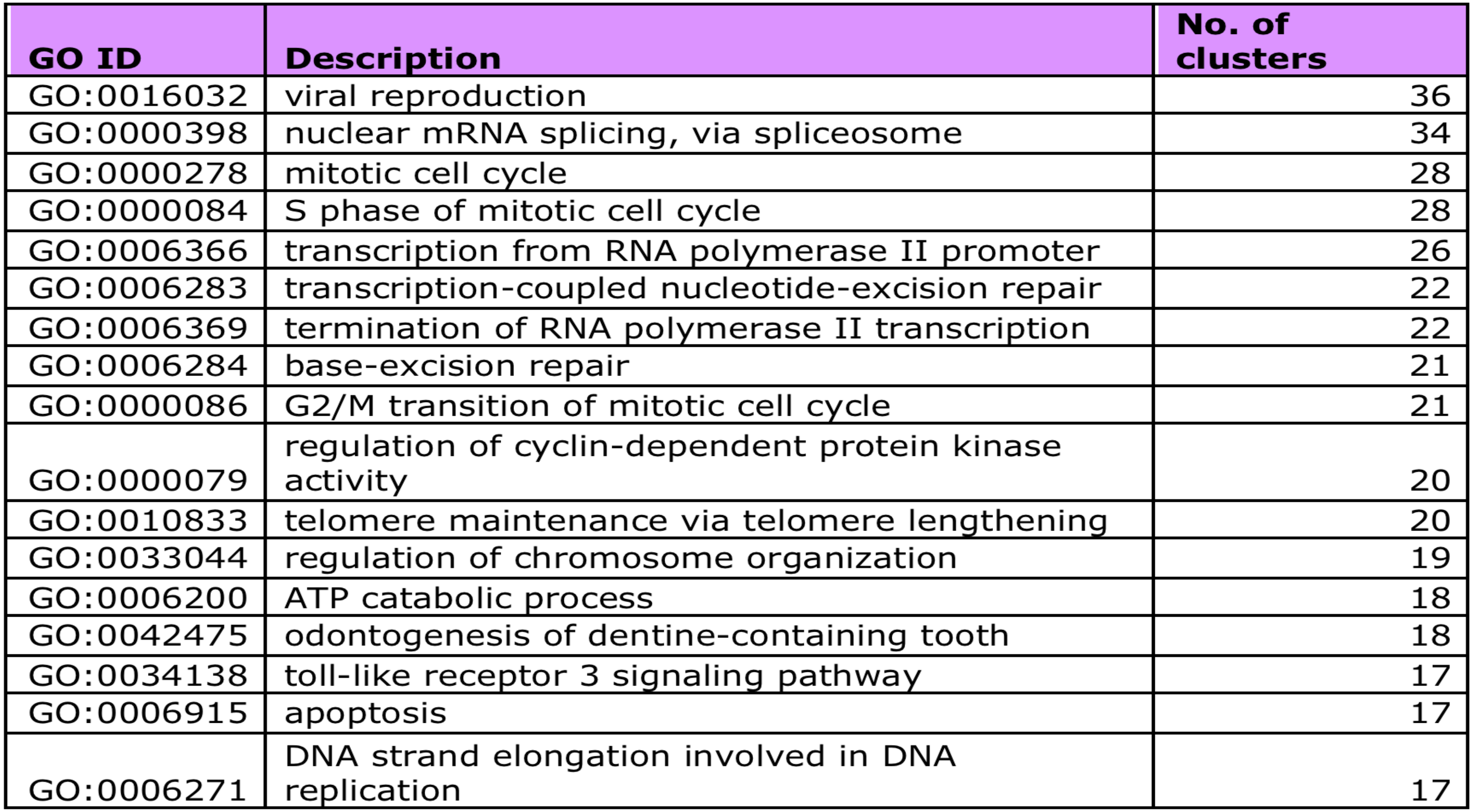
Goh et al. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics. *Journal of Proteome Research*. 11(3):1571-1581, March 2012

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*. 11(3):1571-1581, March 2012.

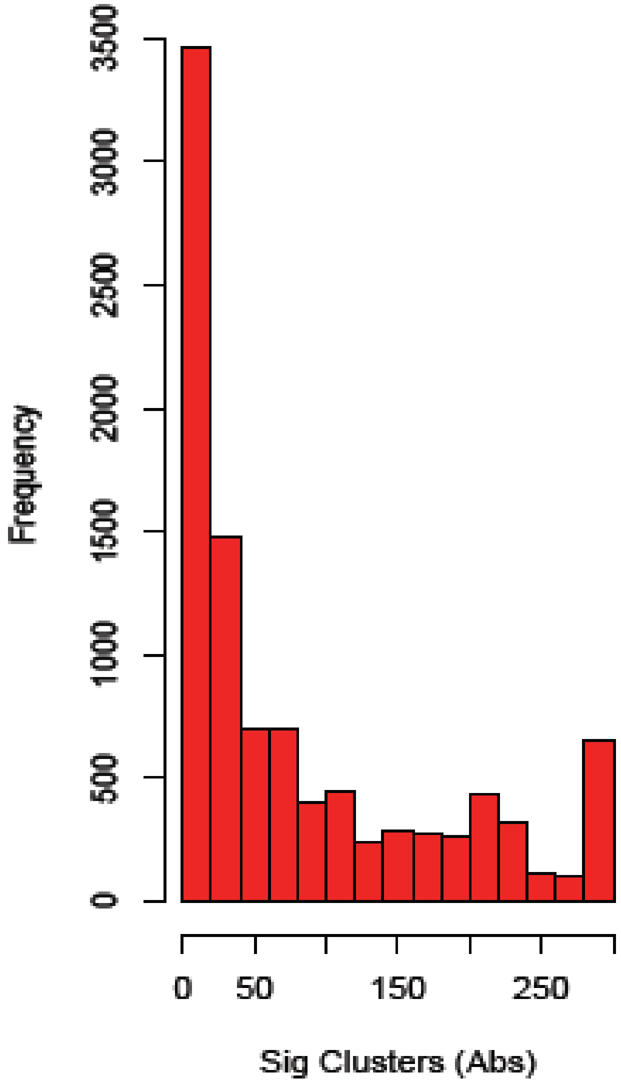
Top-Ranked GO Terms



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

Goh et al**. Enhancing utility of proteomics signature profiling (PSP) with pathway derived subnets (PDSs), performance analysis and specialized ontologies**. *BMC Genomcs, to appear.*

False Positive Rate Analysis



- **Divide 7 poor patients into 2 groups**
  - Significant complexes produced by PSP here are false positives
- **Repeat many times to get dull distribution**
  - Median = 40, mode = 6
- **Cf. 523 complexes in CORUM (size ≥4) used in PSP. At p ≤ 5%, 523 * 5% ≈ 27 false positives expected**

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)

Goh et al**. Enhancing utility of proteomics signature profiling (PSP) with pathway derived subnets (PDSs), performance analysis and specialized ontologies**. *BMC Genomcs, 14:35, 2013.*

Pathway-Derived Subnets (PDS)

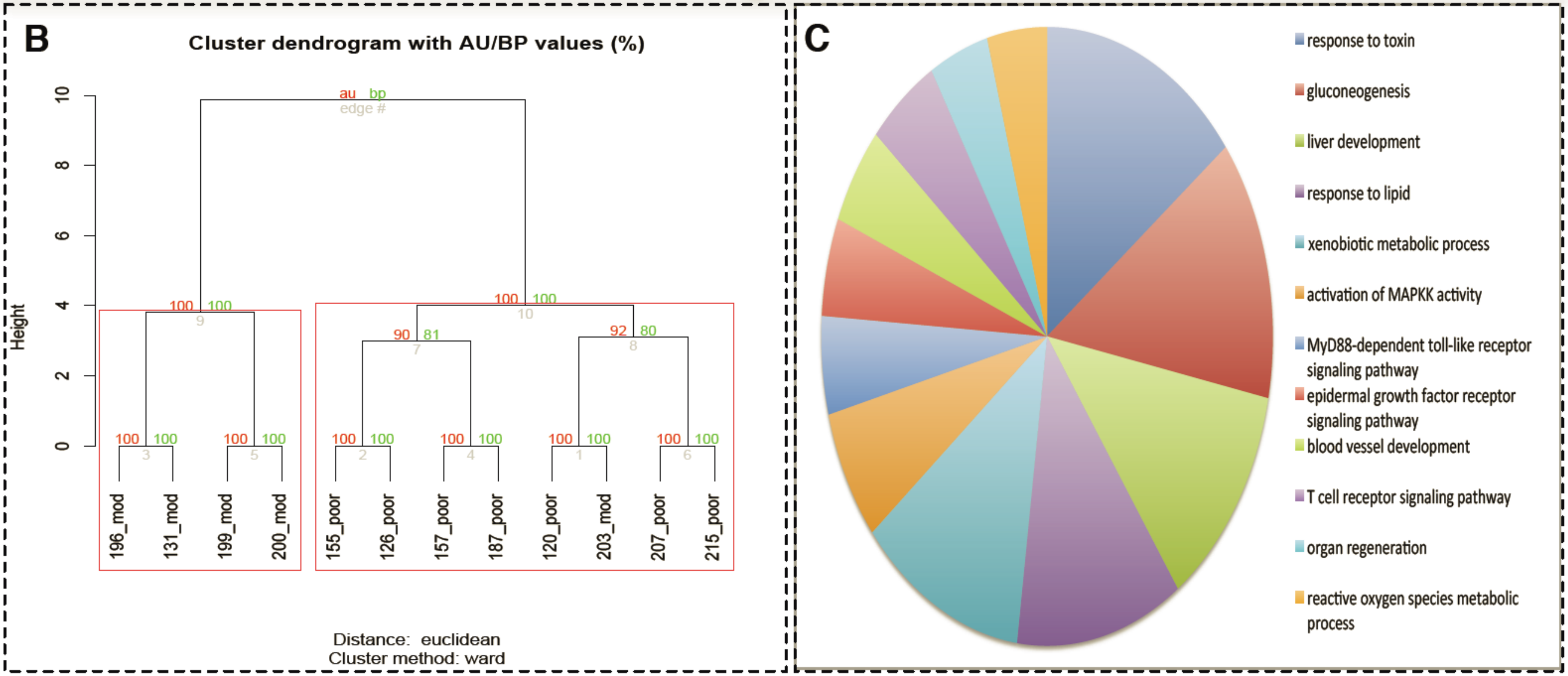
- **Identify the set S_i of proteins detected in more than 50% of samples having phenotype P_i**
  - Do this for each phenotype $P_1, \ldots, P_k$
- **Overlay $\cup_i S_i$ to pathways**
- **Remove nodes not covered by $\cup_i S_i$**

  $\Rightarrow$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

Goh et al. **Enhancing utility of proteomics signature profiling (PSP) with pathway derived subnets (PDSs), performance analysis and specialized ontologies**. *BMC Genomcs, 14:35, 2013.*

PDS consistently segregates mod vs poor patients



Source: Wilson Goh

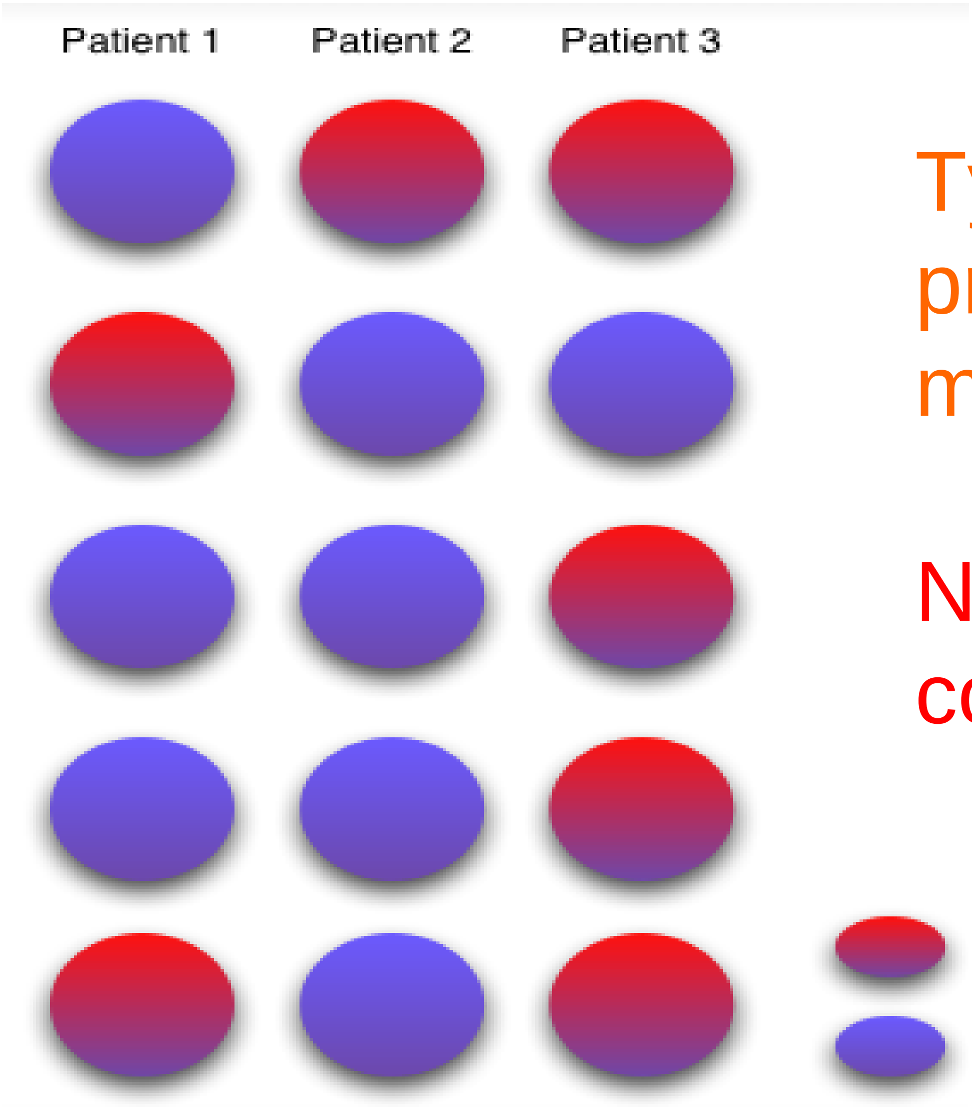
What have we learned?

- 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

Improving Coverage in Proteomic Profile Analysis

NUS
National University
of Singapore

Typical proteomic profiling misses many proteins



Need to improve coverage!

Image credit: Wilson Goh

FCS

- **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 containing many high-confidence proteins**

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

- **Shortcoming: Cliques are too strict**

⇒ **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, 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 …

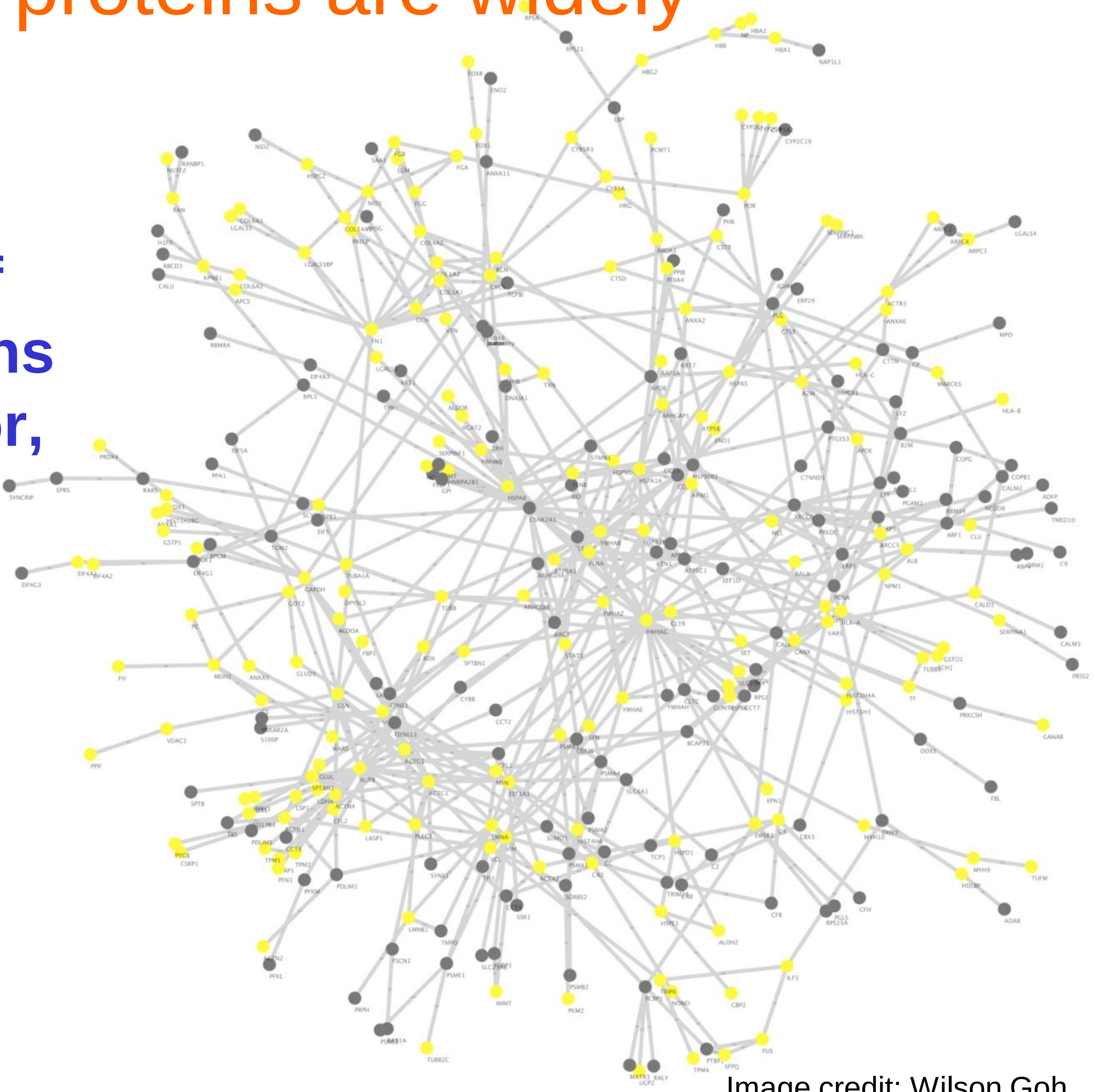
Background

- **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

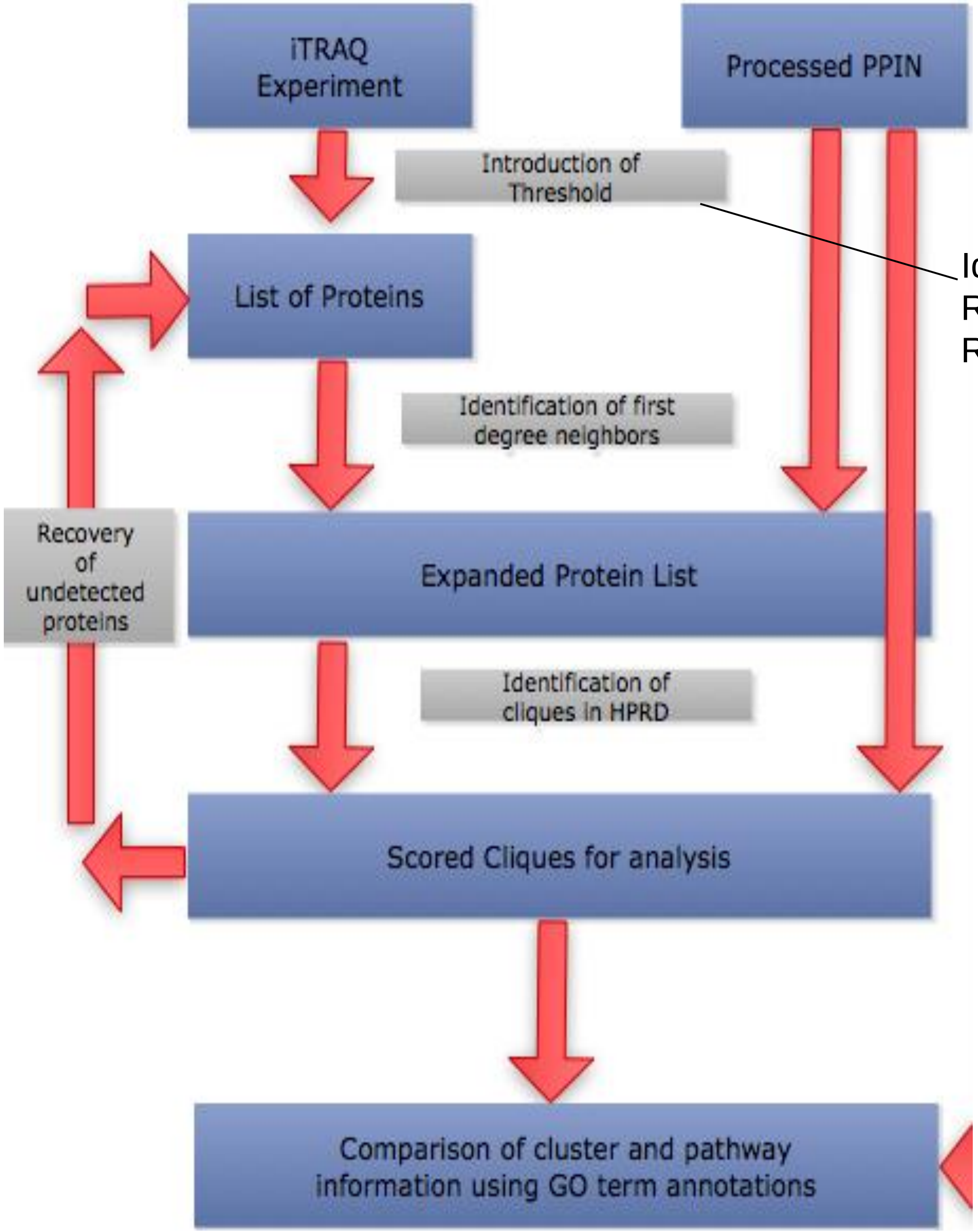
- **In the subnet of reported proteins in mod and poor, poor and mod genes are well mixed**



● Mod and Poor

● Poor only

Image credit: Wilson Goh


iTRAQ Experiment
Processed PPIN
Introduction of Threshold
List of Proteins
Identification of first degree neighbors
Recovery of undetected proteins
Expanded Protein List
Identification of cliques in HPRD
Scored Cliques for analysis
Comparison of cluster and pathway information using GO term annotations

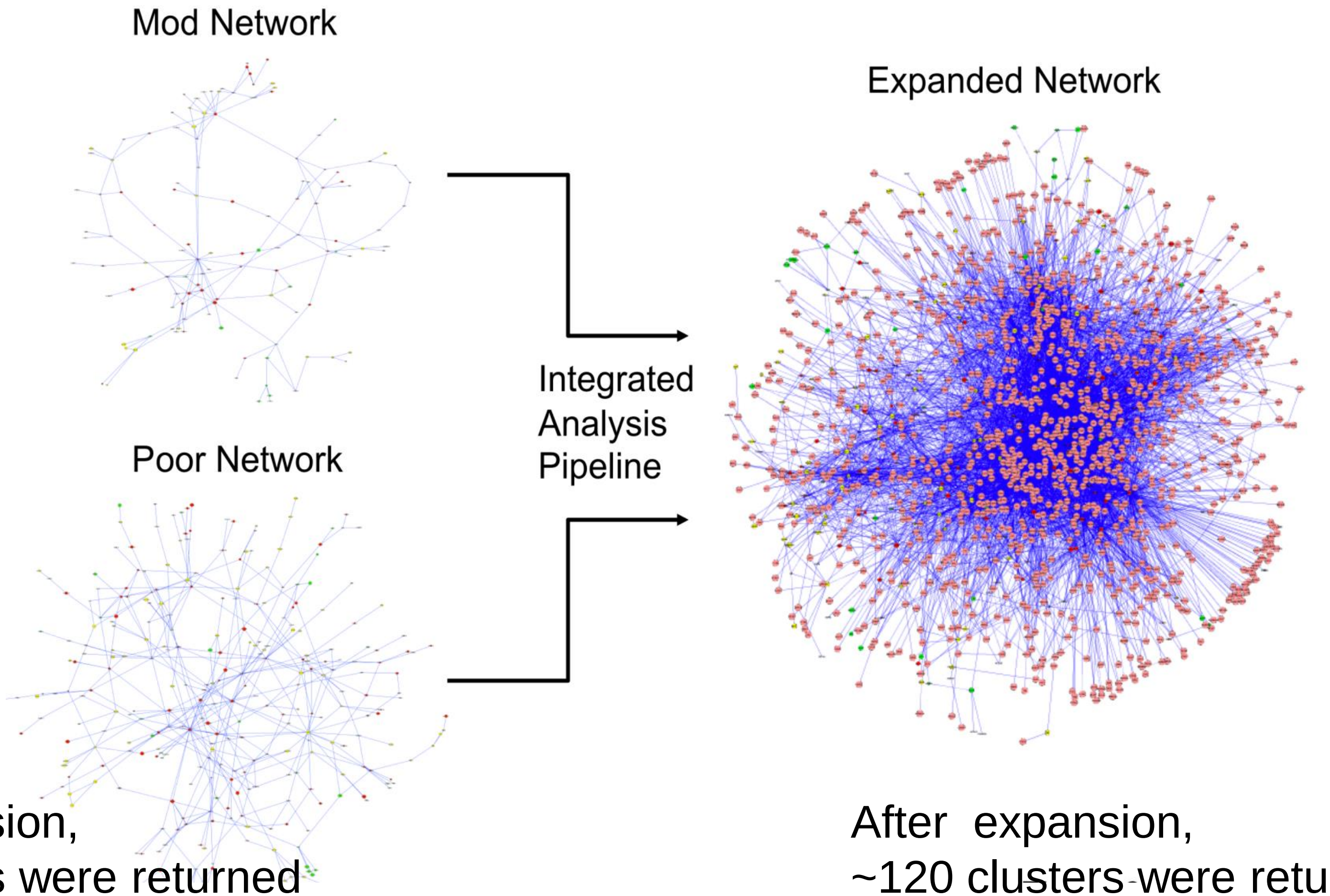
Identify the “seeds”
Ratio < 0.8 and > 1.25 for Mod (min 3 patients)
Ratio < 0.8 and > 1.25 for Poor (min 4 patients)

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 ($p < 0.05$)**

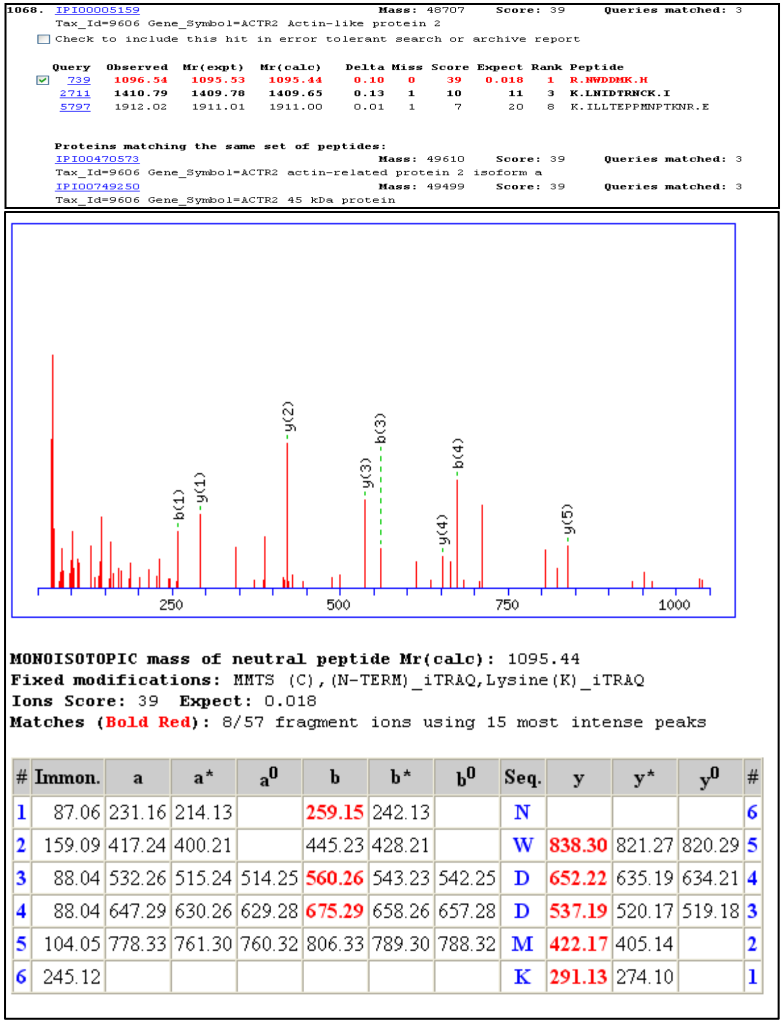
⇒ In silico expansion verified

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

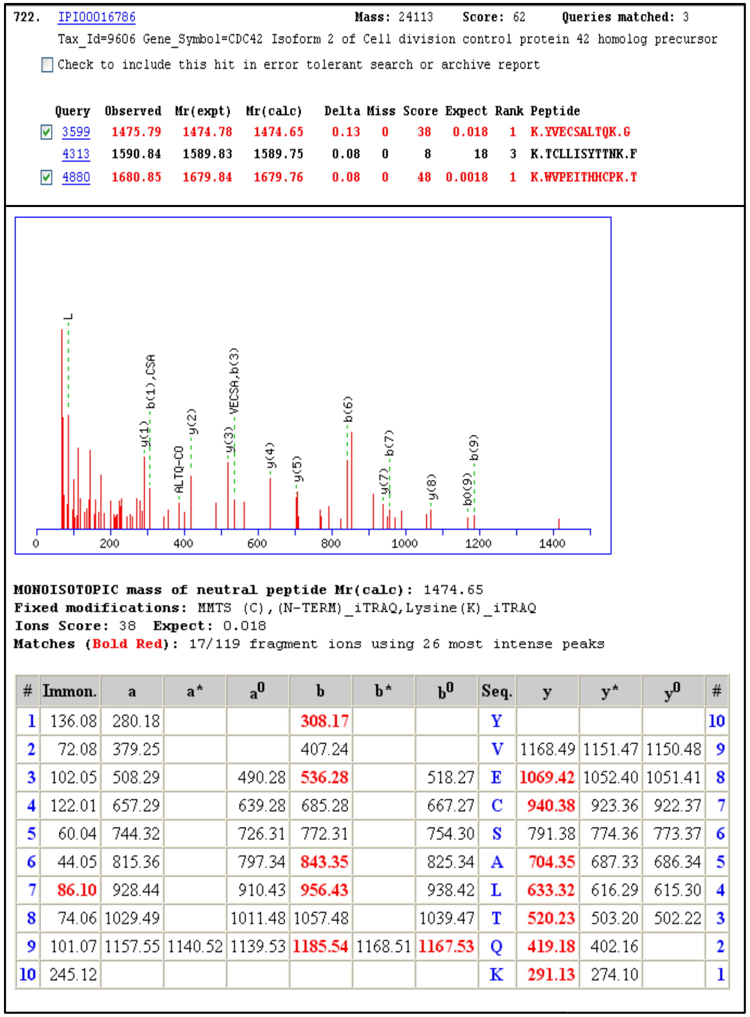


```
1068.  IPI00005159                         Mass: 48707    Score: 39     Queries matched: 3
       Tax_Id=9606 Gene_Symbol=ACTR2 Actin-like protein 2
     □ Check to include this hit in error tolerant search or archive report

       Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score Expect Rank Peptide
     ☑  739   1096.54   1095.53   1095.44    0.10    0   39   0.018    1  R.NWDDMK.H
       2711   1410.79   1409.78   1409.65    0.13    1   10      11    3  K.LNIDTRNCK.I
       5797   1912.02   1911.01   1911.00    0.01    1    7      20    8  K.ILLTEPPMNPTKNR.E


       Proteins matching the same set of peptides:
       IPI00470573                         Mass: 49610    Score: 39     Queries matched: 3
       Tax_Id=9606 Gene_Symbol=ACTR2 actin-related protein 2 isoform a
       IPI00749250                         Mass: 49499    Score: 39     Queries matched: 3
       Tax_Id=9606 Gene_Symbol=ACTR2 45 kDa protein
```

```
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



```
722.  IPI00016786                     Mass: 24113    Score: 62     Queries matched: 3
      Tax_Id=9606 Gene_Symbol=CDC42 Isoform 2 of Cell division control protein 42 homolog precursor
    □ Check to include this hit in error tolerant search or archive report

      Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score Expect Rank Peptide
    ☑ 3599   1475.79   1474.78   1474.65    0.13    0   38   0.018    1  K.YVECSALTQK.G
      4313   1590.84   1589.83   1589.75    0.08    0    8      18    3  K.TCLLISYTTNK.F
    ☑ 4880   1680.85   1679.84   1679.76    0.08    0   48  0.0018    1  K.WVPEITHHCPK.T
```

```
MONOISOTOPIC mass of neutral peptide Mr(calc): 1474.65
Fixed modifications: MMTS (C),(N-TERM)_iTRAQ,Lysine(K)_iTRAQ
Ions Score: 38  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

Another Experiment: Comparison

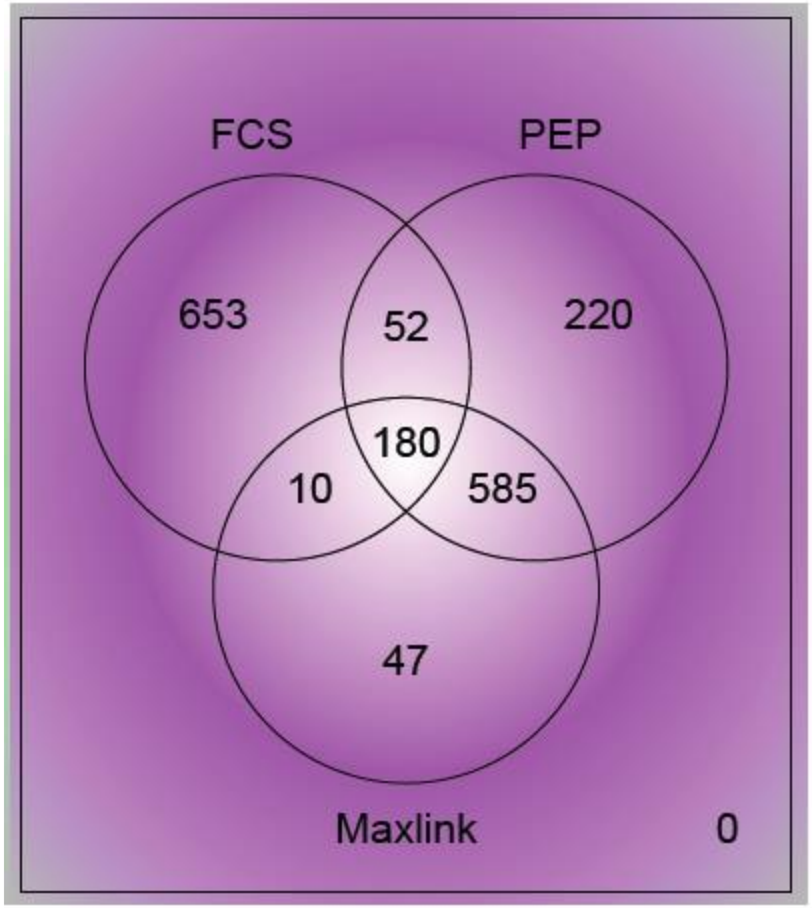
- **Valporic acid (VPA)-treated mice vs control**
  - VPA or vehicle injected every 12 hours into postnatal day-56 adult mice for 2 days
  - Role of VPA in epigenetic remodeling

- **MS was scanned against IPI rat db in round #1**
  - 291 proteins identified
- **MS was scanned against UniProtkb in round #2**
  - 498 additional proteins identified

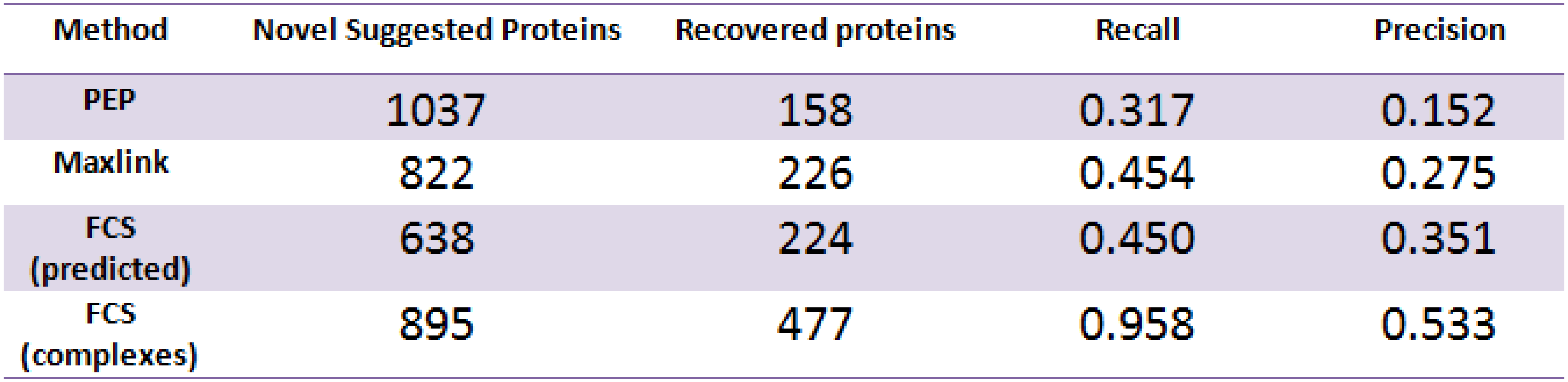
- **All recovery methods ran on round #1 data and the recovered proteins checked against round #2**

Moderate level of agreement of reported proteins between various recovery methods

FCS (Real Complexes)



Performance Comparison



Method	Novel Suggested Proteins	Recovered proteins	Recall	Precision
PEP	1037	158	0.317	0.152
Maxlink	822	226	0.454	0.275
FCS (predicted)	638	224	0.450	0.351
FCS (complexes)	895	477	0.958	0.533

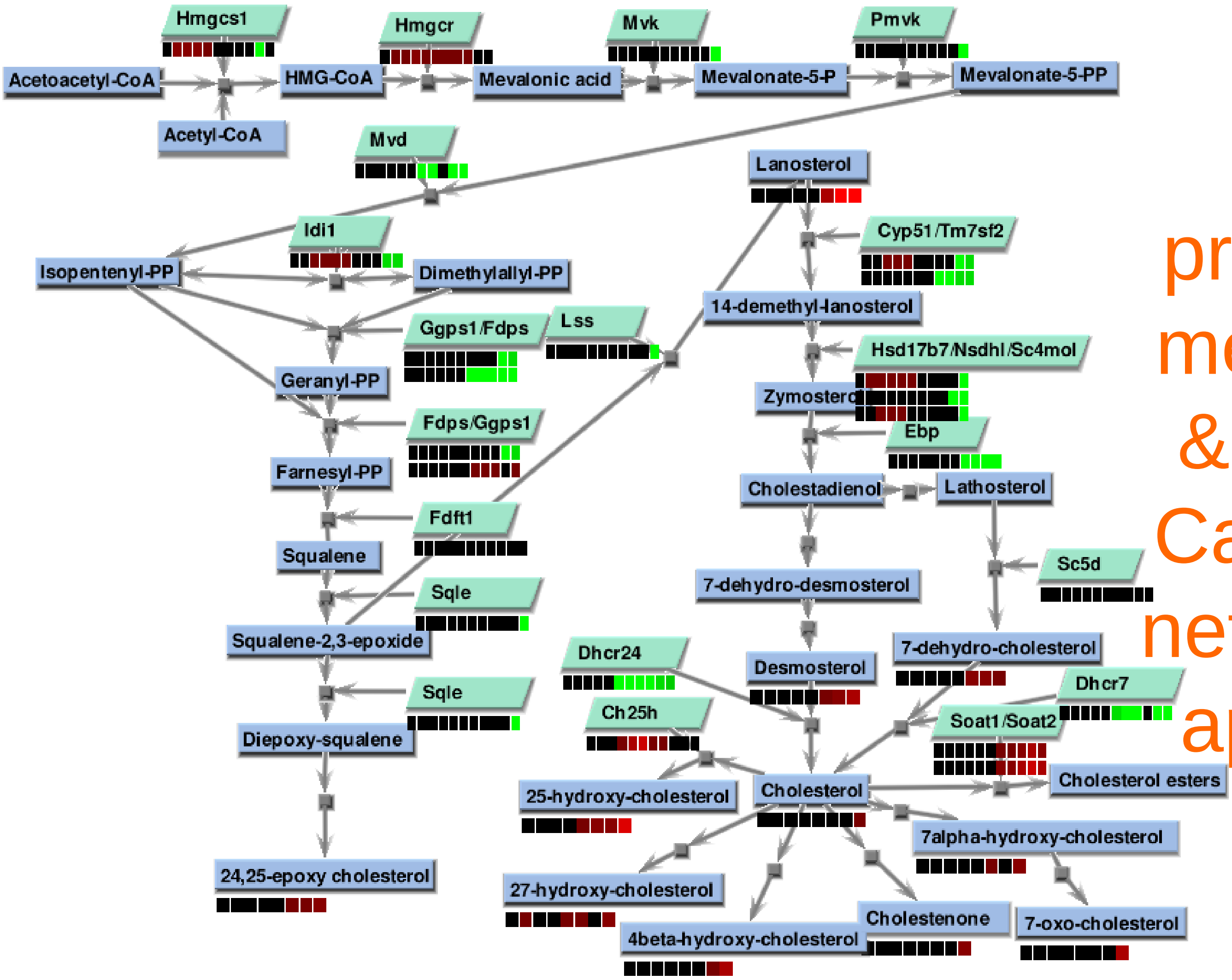
- **Looks like running FCS on real complexes is able to recover more proteins and more accurately**

Must Read

- Steen & Mann. **The ABC's and XYZ's of peptide sequencing.** *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004
- Käll & Vitek. **Computational mass spectrometry–based proteomics**. *PLoS Comput Biol* , 7(12): e1002277, 2011
- Cottrell. Protein identification using MS/MS data. Journal of Proteomics, 74:1842-1851, 2011
- Goh et al. **How advancement in biological network analysis methods empowers proteomics**. *Proteomics*, 12(4-5):550-563, 2012

Good to Read

- [PSP] Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. 11(3):1571-1581, 2012
- [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**. *Int J Bioinform Res and App, 8(3/4):155-170, 2012*
- Frank, et al. **De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry.** *J. Proteome Res*. 6:114-123, 2007



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

Acknowledgements

- **The slides on peptide identification were adapted from those given to me by A/P Leong Hon Wai**

- **A lot of the slides on PSP, PDS, and PEP came from the work of Wilson Goh**

Leong Hon Wai

Wilson Goh