Proteomics Informatics (BMSC-GA 4437)

Instructor

David Fenyö

Contact information

David@FenyoLab.org

http://fenyolab.org/presentations/Proteomics_Informatics_2013/

computational proteomics

Laboratory of Computational Proteomics

Proteomics Informatics Spring 2013

Overview

Research

Members

Publications

Presentations

Tools

Contact

Proteomics Informatics Spring 2013

Instructor: David Fenyö, Associate Professor
Contact information: David@FenyoLab.org

Week 1 Overview of proteomics (1/29/2013 at 4 pm in TRB 718)

Reading list

- M.A. Gillette, S.A. Carr, ***"Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry"***, Nature Methods 10 (2013) 28–34.
- A. Bensimon, A.J.R. Heck R. Aebersold, ***"Mass Spectrometry–Based Proteomics and Network Biology"***, Annual Review of Biochemistry 81 (2012) 379-405.

Week 2 Overview of mass spectrometry (2/5/2013 at 4 pm in TRB 718)

Reading list

- Beavis, R.C. & Chait, B.T. ***"Matrix-assisted laser desorption ionization mass-spectrometry of proteins"*** Meth. Enzymol 270, 519-551 (1996).
- Banks, J.F. & Whitehouse, C.M. ***"Electrospray ionization mass spectrometry"*** Meth. Enzymol 270, 486-519 (1996).
- Chalkley, R. ***"Instrumentation for LC-MS/MS in proteomics"*** Methods Mol. Biol 658, 47-60 (2010).

Week 3 Analysis of mass spectra: signal processing, peak finding, and isotope clusters (2/12/2013 at 4 pm in TRB 119)

Reading list

- Zhang, J., Gonzalez, E., Hestilow, T., Haskins, W. & Huang, Y. Review of peak

Center for Health Informatics and Bioinformatics

NYU Langone Medical Center

National Institutes of Health

National Cancer Institute

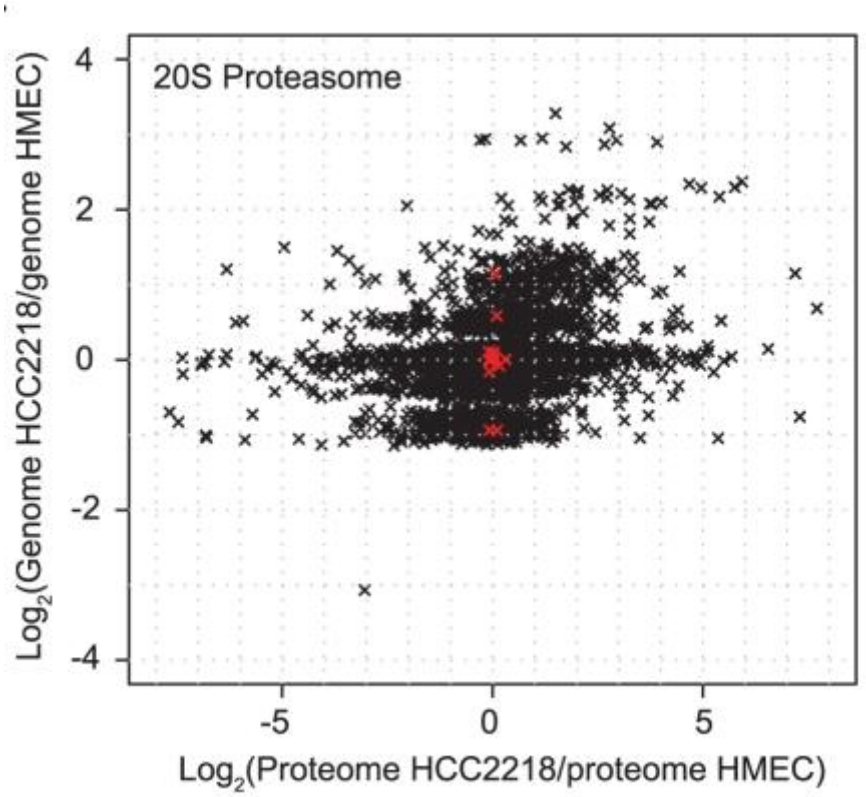
Proteomics Informatics – Learning Objectives

Be able analyze a proteomics data set and understand the limitations of the results.

Proteomics Informatics – Overview of Proteomics (Week 1)

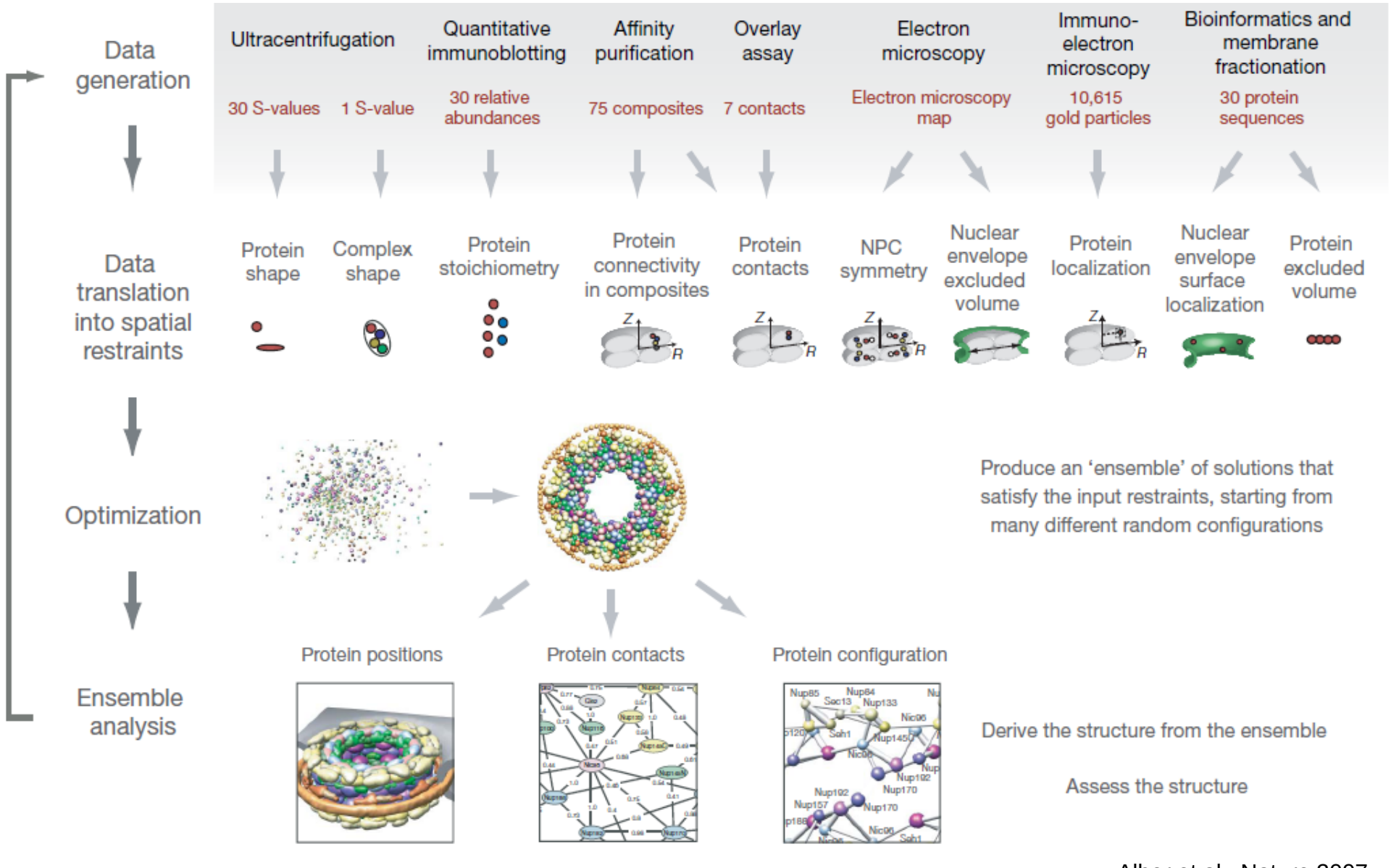
- Why proteomics?
- Bioinformatics
- Overview of the course

Motivating Example: Protein Regulation



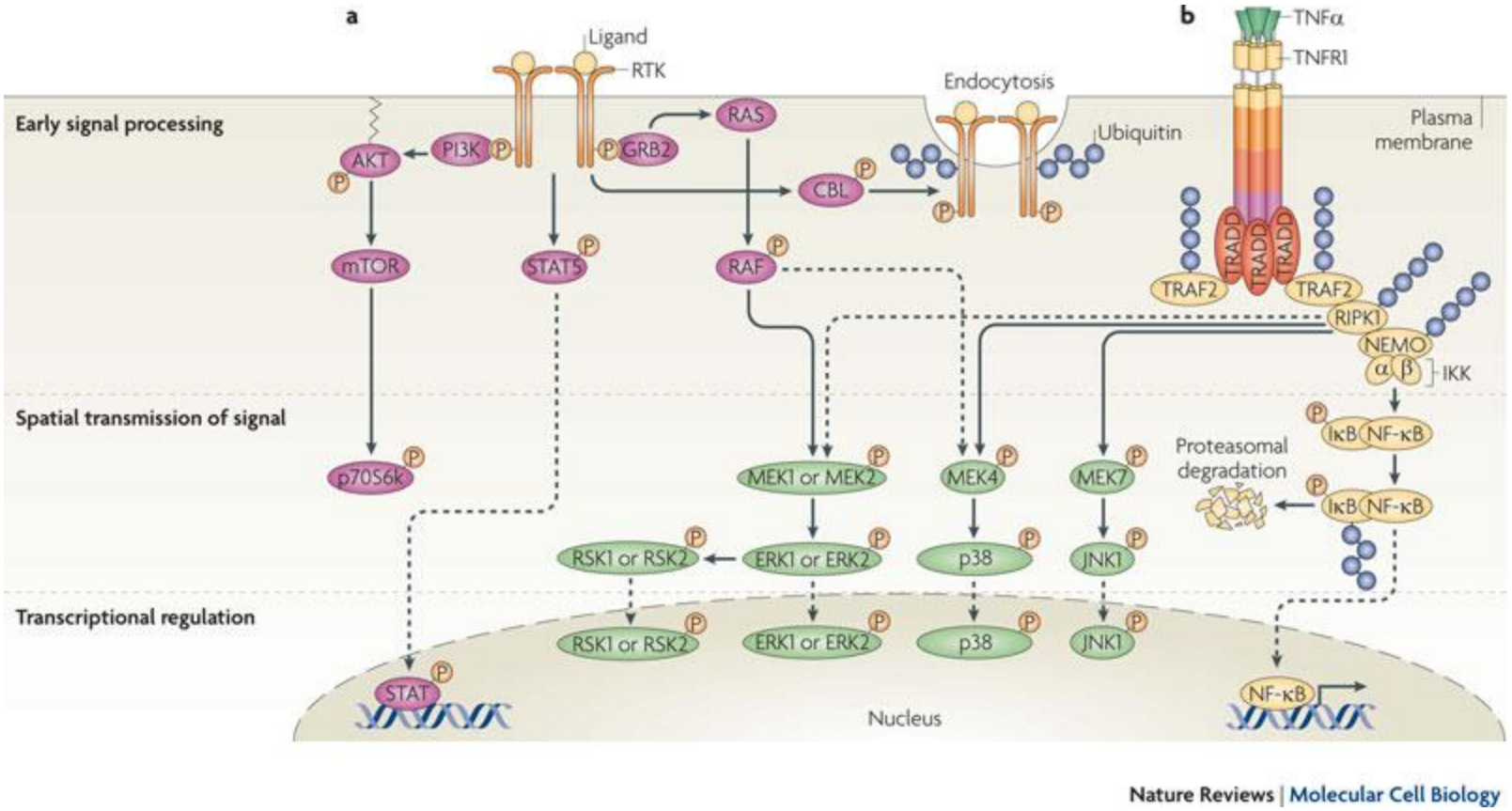
Geiger et al., “Proteomic changes resulting from gene copy number variations in cancer cells”, PLoS Genet. 2010 Sep 2;6(9). pii: e1001090.

Motivating Example: Protein Complexes



Alber et al., Nature 2007

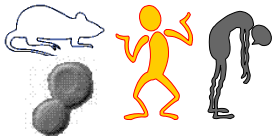
Motivating Example: Signaling



Choudhary & Mann, Nature Reviews Molecular Cell Biology 2010

Bioinformatics

Biological System



↓ Experimental Design

Samples

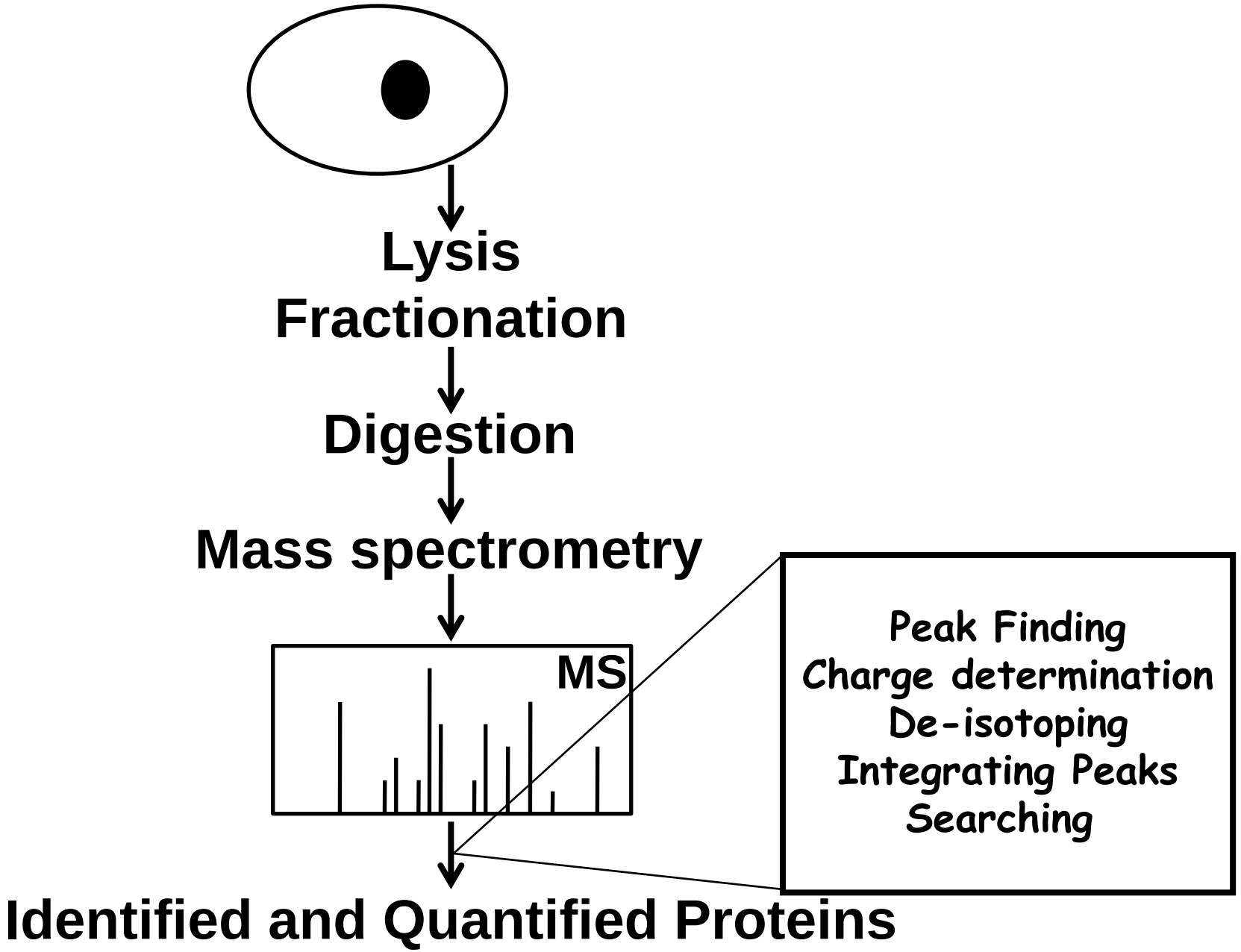
↓ Measurements

Raw Data

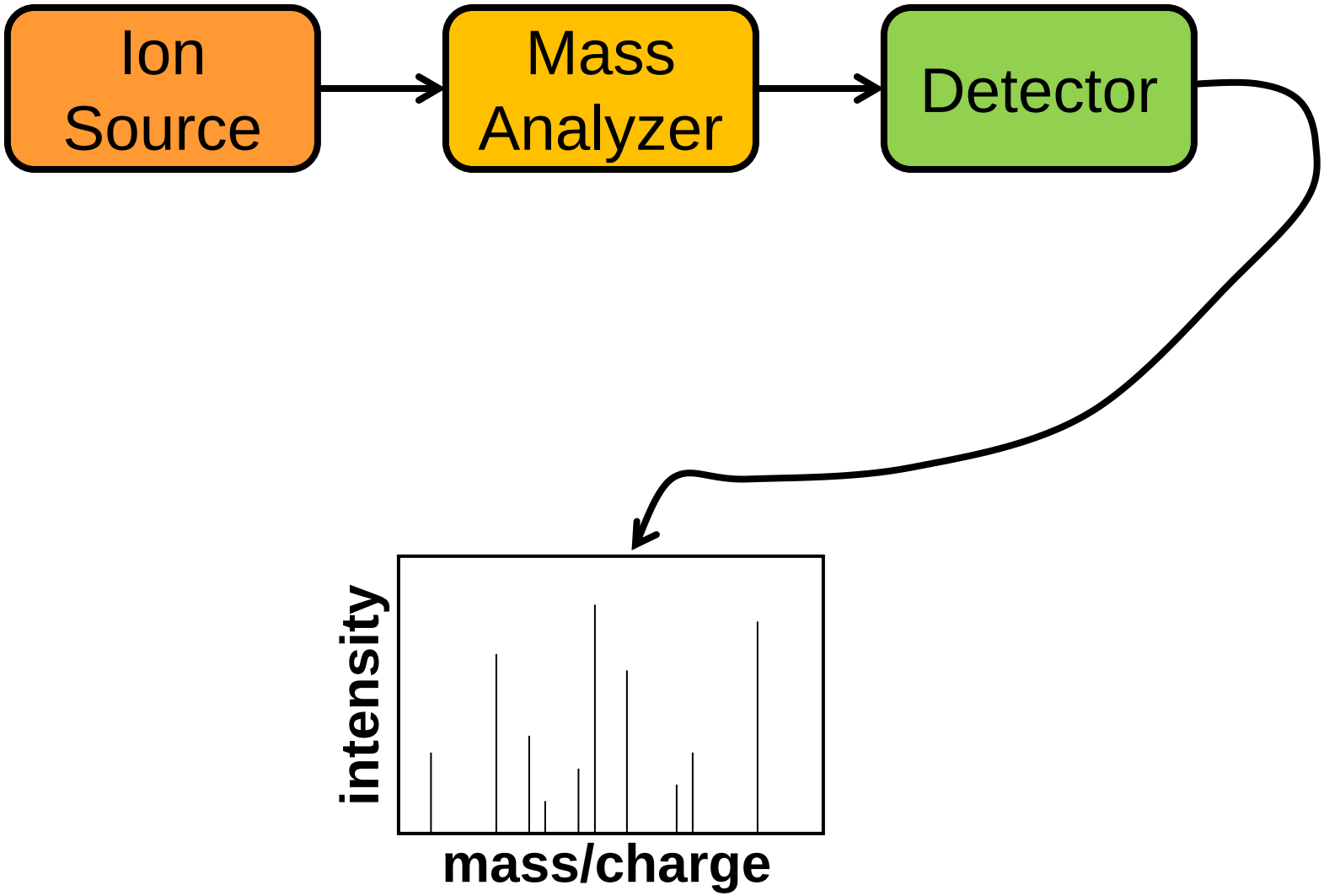
↓ Data Analysis

Information

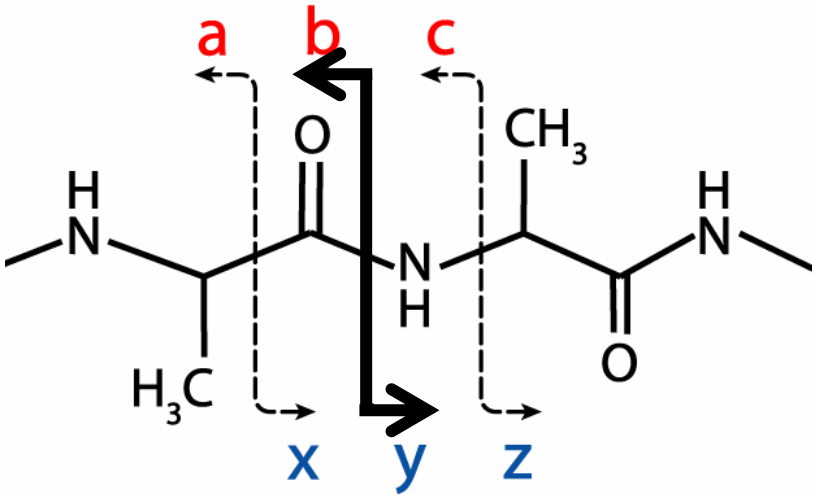
Mass Spectrometry Based Proteomics

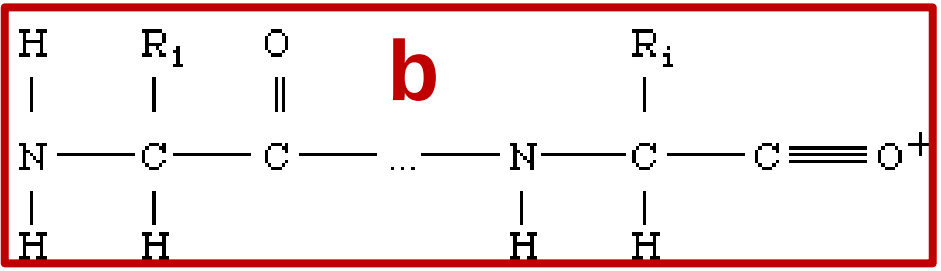


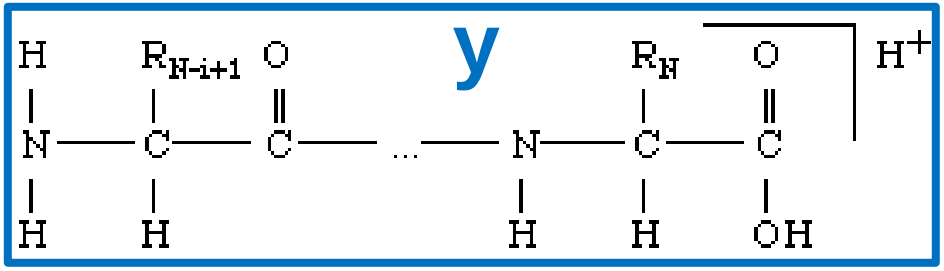
Proteomics Informatics – Overview of Mass spectrometry (Week 2)



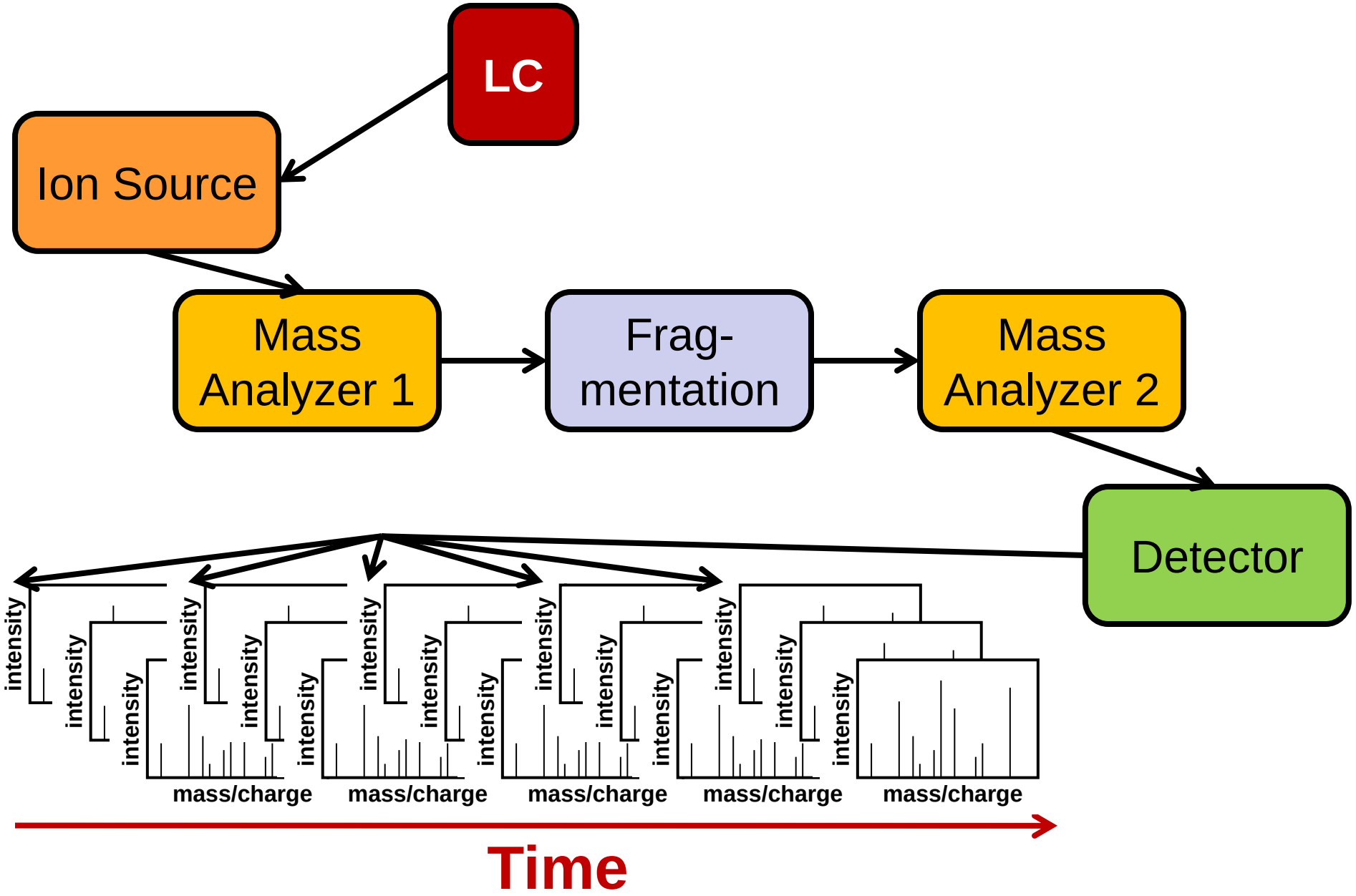
Proteomics Informatics – Overview of Mass spectrometry (Week 2)



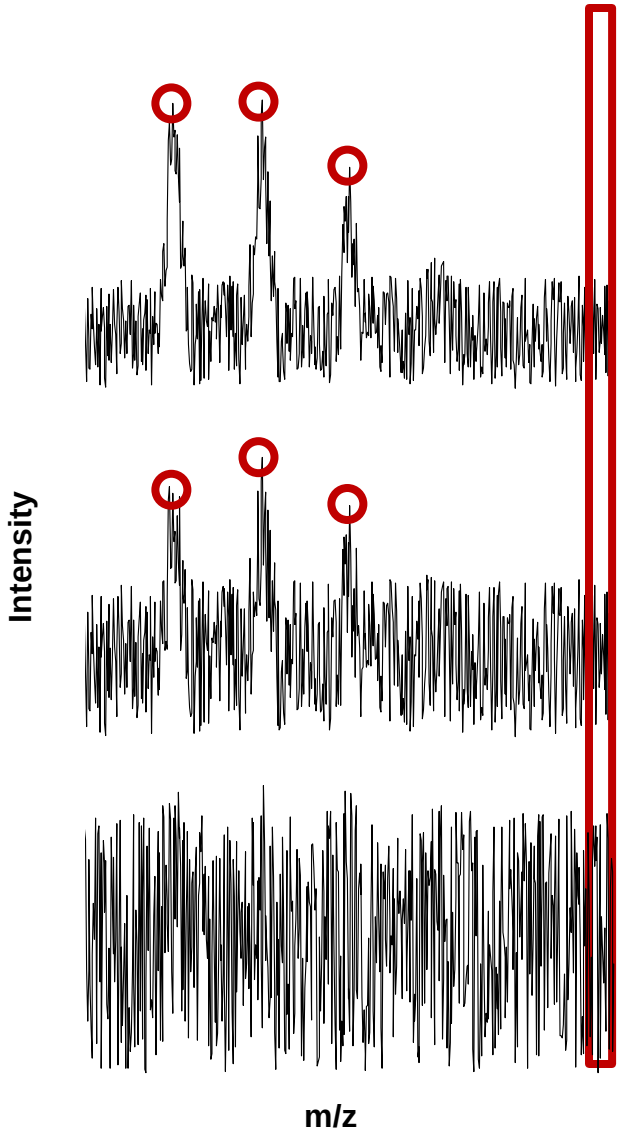




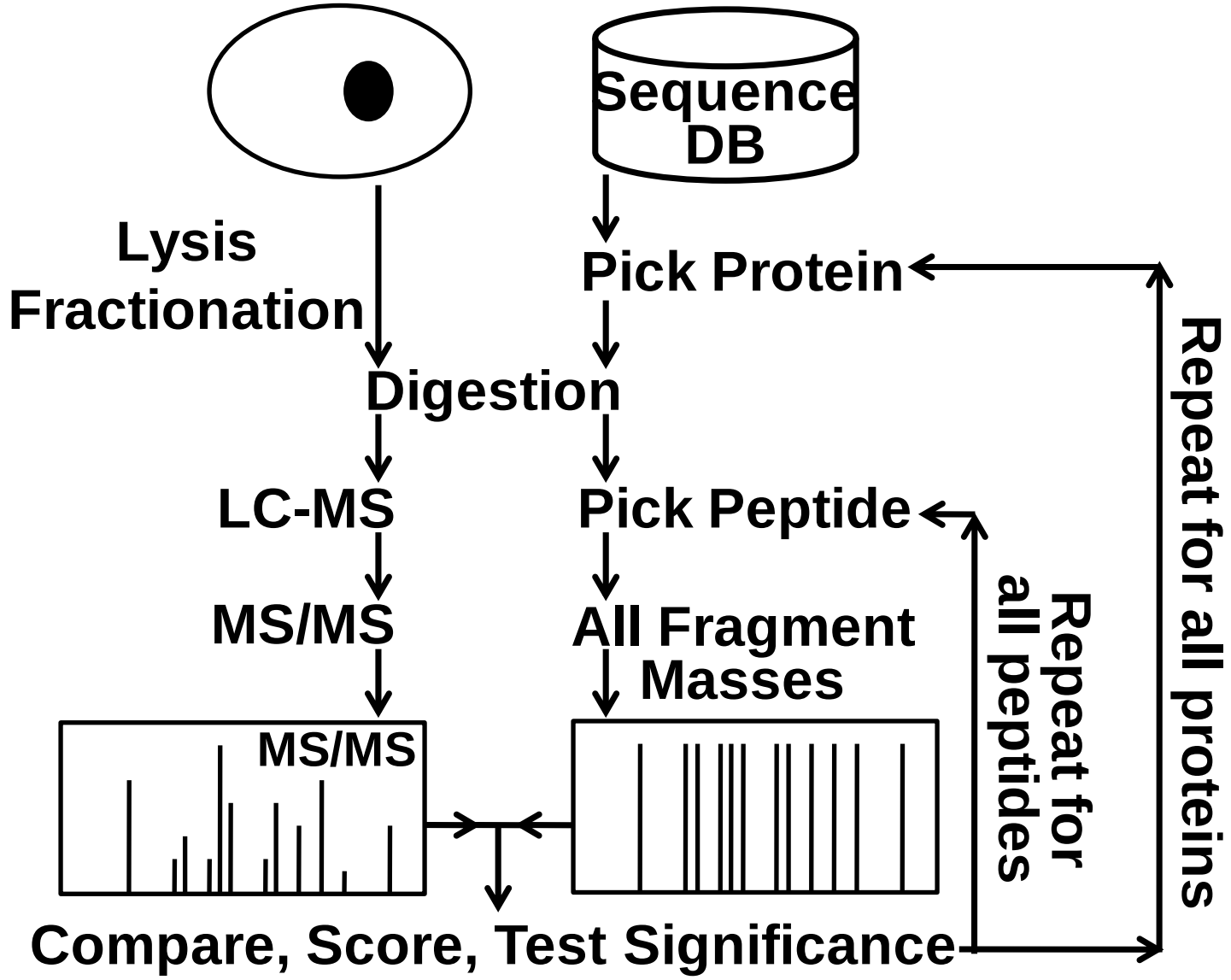
Proteomics Informatics – Overview of Mass spectrometry (Week 2)



Proteomics Informatics – Analysis of mass spectra: signal processing, peak finding, and isotope clusters (Week 3)



Proteomics Informatics – Protein identification I: searching protein sequence collections and significance testing (Week 4)



Proteomics Informatics – Protein identification II: search engines and protein sequence databases (Week 5)

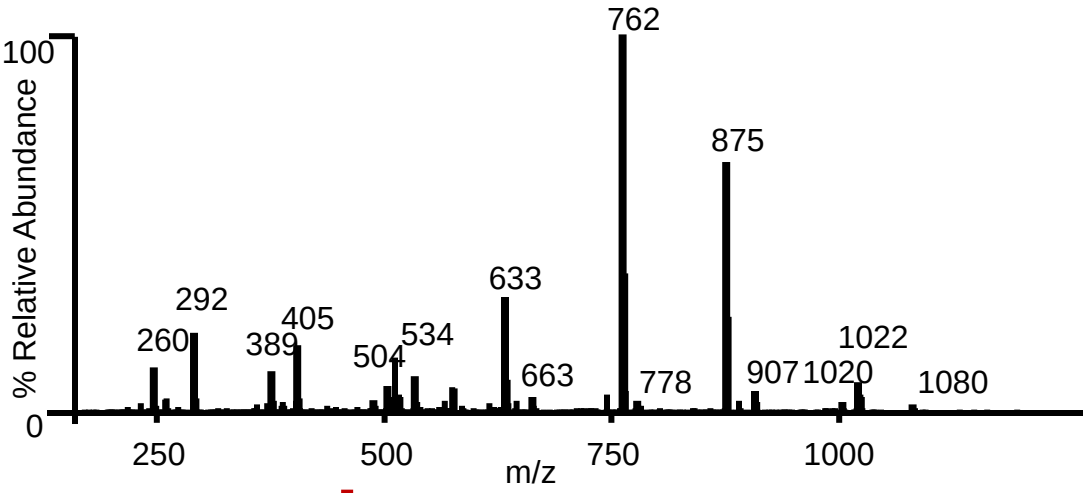
rank	log(e)	log(I)	%/%	#	total	Mr	Accession
1	-673.5	5.10	59/77	59	190	122.8	ENSP00000323315 gpmDB | psyt | snap [2/877] homo (0/9) protein ABL1, c-abl oncogene 1, non-receptor tyrosine kinase [Source: HGNC 76] IPR015015 *(×2)* F-actin binding IPR000719 *(×2)* Prot kinase cat dom IPR000980 *(×8)* SH2 IPR011511 SH3 2 IPR001452 *(×3)* SH3 domain IPR001245 *(×6)* Ser-Thr/Tyr kinase cat dom IPR002290 Ser/Thr dual-sp kinase dom IPR020635 Tyr kinase cat dom
2	-531.9	5.07	43/63	45	155	143.1	ENSP00000314499 gpmDB | psyt | snap [8/1446] homo (0/2) protein GAK, cyclin G associated kinase [Source: HGNC 4113] IPR001623 DnaJ N IPR014019 Phosphatase tensin-typ IPR000719 Prot kinase cat dom IPR001245 Ser-Thr/Tyr kinase cat dom IPR002290 Ser/Thr dual-sp kinase dom IPR014020 Tensin phosphatase C2-dom IPR020635 Tyr kinase cat dom
3	-508.8	5.18	40/55	45	178	142.7	ENSP00000303507 gpmDB | psyt | snap [0/1011] homo (1/12) protein BCR, breakpoint cluster region [Source: HGNC 1014] IPR015123 Bcr-Abl oncoprot oligo IPR000008 *(×2)* C2 Ca-dep IPR018029 C2 membr targeting IPR000219 *(×3)* DH-domain IPR001849 *(×3)* Pleckstrin homology IPR000198 *(×3)* RhoGAP dom
4	-471.1	4.76	34/48	44	74	181.6	ENSP00000375986 gpmDB | psyt | snap [0/569] homo (4/4) protein MAP3K4, mitogen-activated protein kinase kinase kinase 4 [Source: HGNC 6856] IPR000719 Prot kinase cat dom IPR002290 Ser/Thr kinase dom IPR020635 Tyr kinase cat dom

1/11

Proteomics Informatics – Protein identification III: de novo sequencing (Week 6)

Amino acid masses

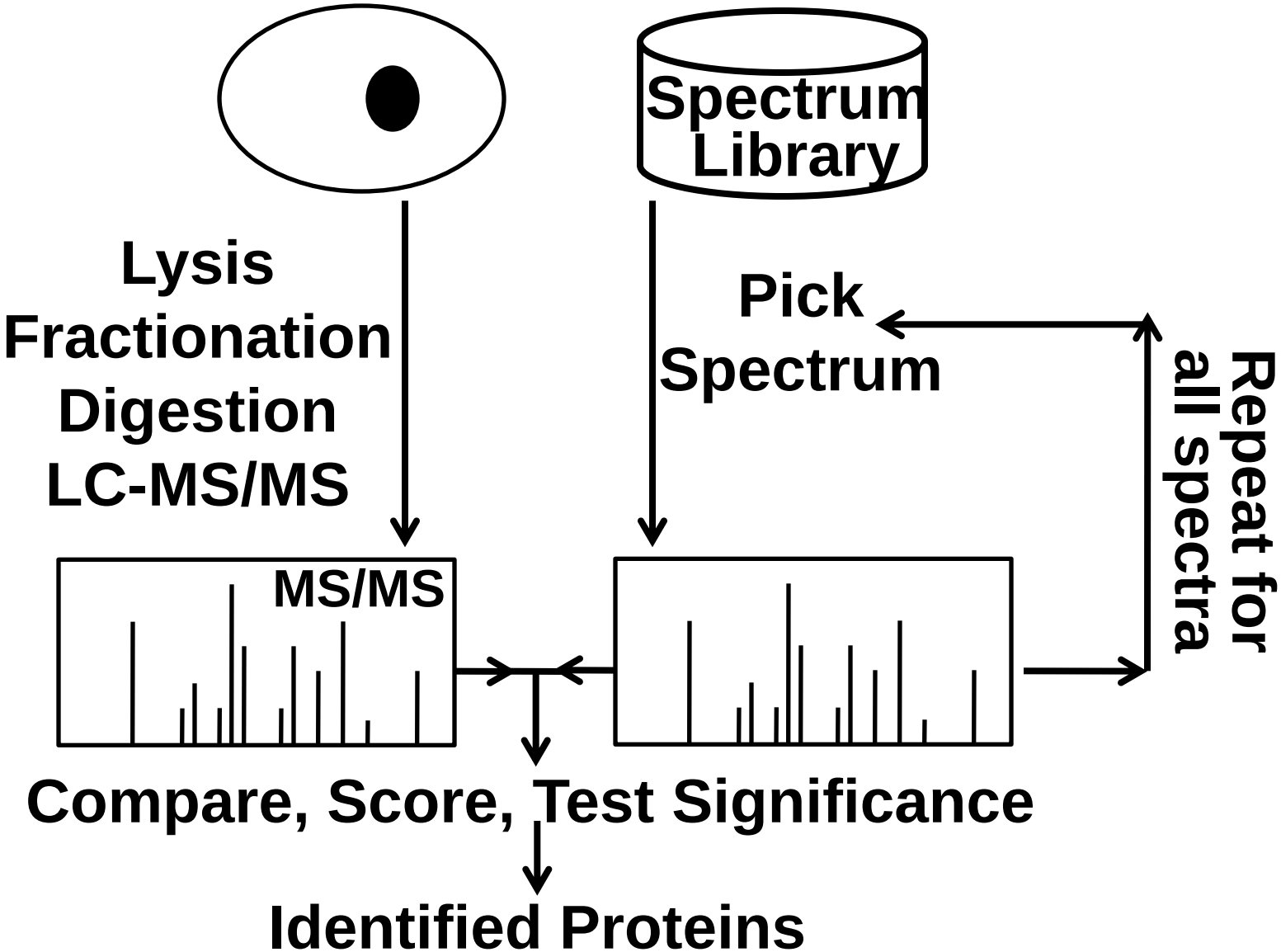
1-letter code	3-letter code	Chemical formula	Monoisotopic	Average
A	Ala	C_3H_5ON	71.0371	71.0788
R	Arg	$C_6H_{12}ON_4$	156.101	156.188
N	Asn	$C_4H_6O_2N_2$	114.043	114.104
D	Asp	$C_4H_5O_3N$	115.027	115.089
C	Cys	C_3H_5ONS	103.009	103.139
E	Glu	$C_5H_7O_3N$	129.043	129.116
Q	Gln	$C_5H_8O_2N_2$	128.059	128.131
G	Gly	C_2H_3ON	57.0215	57.0519
H	His	$C_6H_7ON_3$	137.059	137.141
I	Ile	$C_6H_{11}ON$	113.084	113.159
L	Leu	$C_6H_{11}ON$	113.084	113.159
K	Lys	$C_6H_{12}ON_2$	128.095	128.174
M	Met	C_5H_9ONS	131.04	131.193
F	Phe	C_9H_9ON	147.068	147.177
P	Pro	C_5H_7ON	97.0528	97.1167
S	Ser	$C_3H_5O_2N$	87.032	87.0782
T	Thr	$C_4H_7O_2N$	101.048	101.105
W	Trp	$C_{11}H_{10}ON_2$	186.079	186.213
Y	Tyr	$C_9H_9O_2N$	163.063	163.176
V	Val	C_5H_9ON	99.0684	99.1326



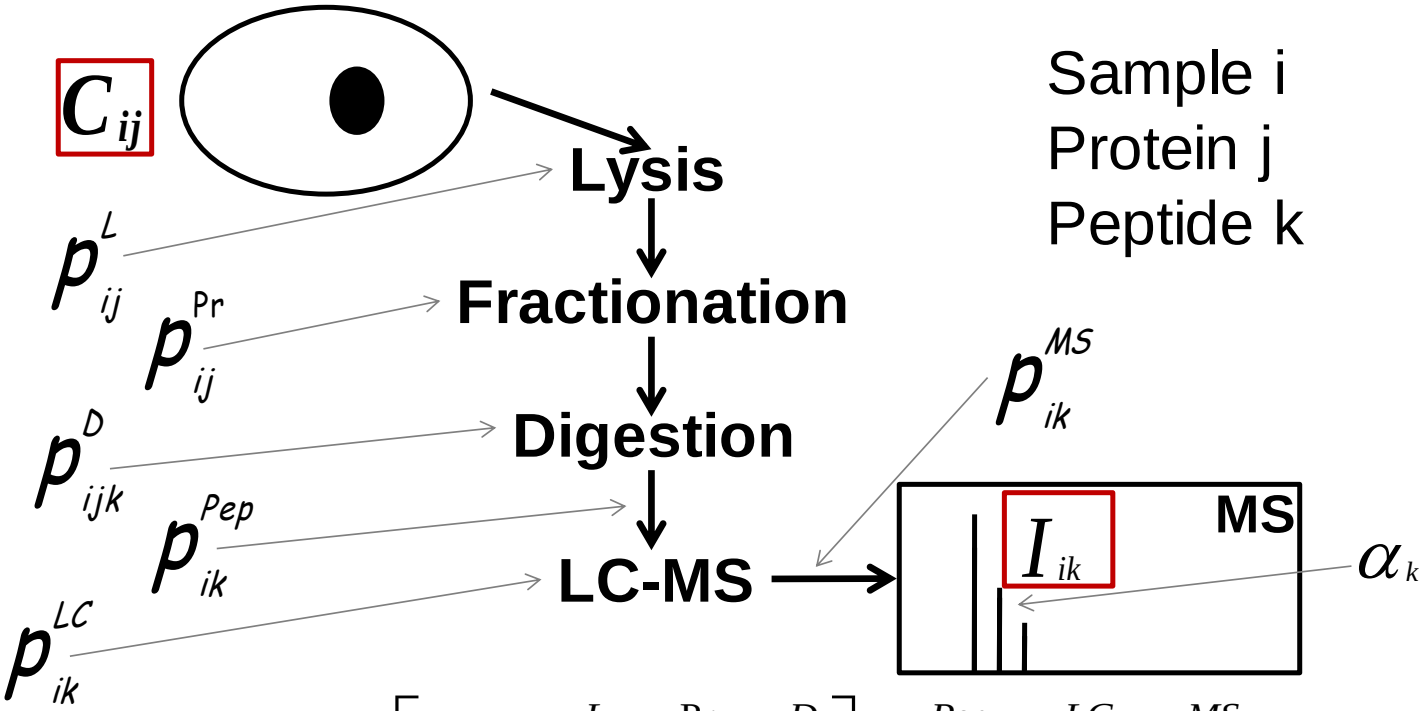
Mass Differences

Sequences consistent with spectrum

Proteomics Informatics – Protein identification IV: spectrum library searching (Week 7)



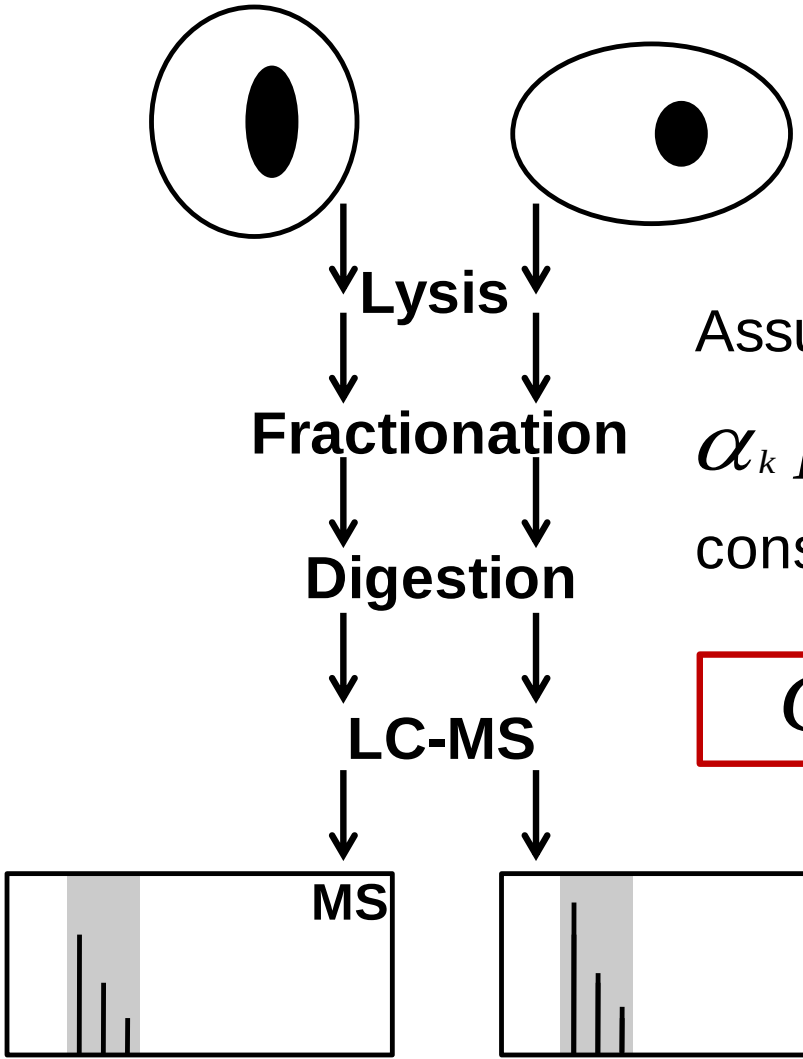
Proteomics Informatics – Protein quantitation I: metabolic labeling (SILAC), chemical labeling, label-free quantitation, spectrum counting (Week 8)



$$I_{ik} = \alpha_k \sum_j \left[C_{ij}\, p^{L}_{ij}\, p^{Pr}_{ij}\, p^{D}_{ijk} \right] p^{Pep}_{ik}\; p^{LC}_{ik}\; p^{MS}_{ik}$$

$$C^{k}_{ij} = \frac{I_{ik}}{\alpha_k\, p^{L}_{ij}\, p^{Pr}_{ij}\, p^{D}_{ijk}\, p^{Pep}_{ik}\; p^{LC}_{ik}\; p^{MS}_{ik}}$$

Proteomics Informatics – Protein quantitation I: metabolic labeling (SILAC), chemical labeling, label-free quantitation, spectrum counting (Week 8)



Sample i
Protein j
Peptide k

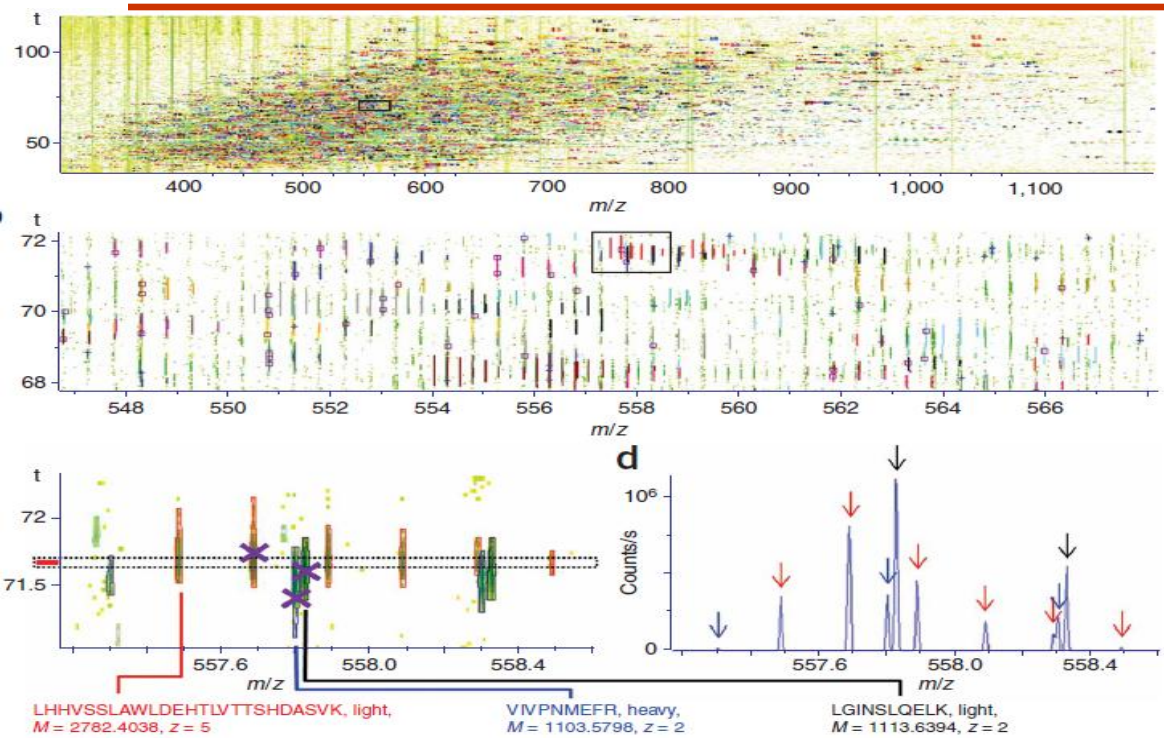
Assumption:

$$\alpha_k p_{ij}^{L} p_{ij}^{\mathrm{Pr}} p_{ijk}^{D} p_{ik}^{Pep} p_{ik}^{LC} p_{ik}^{MS}$$

constant for all samples

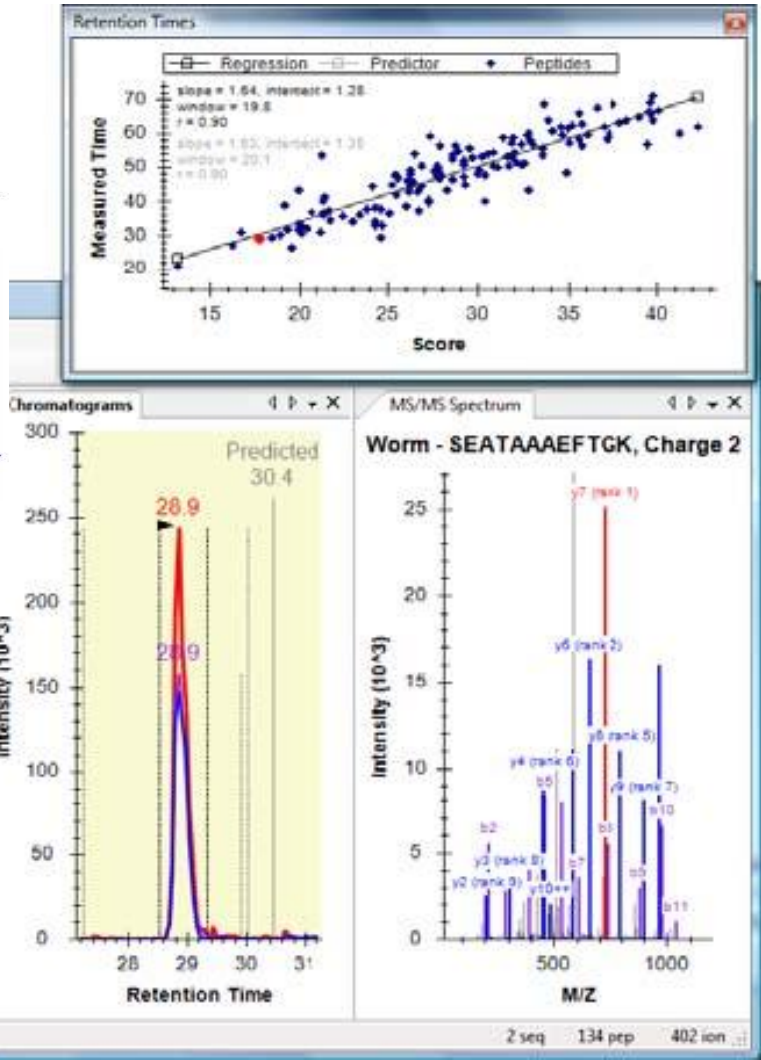
$$C_{i_n j} / C_{i_m j} = I_{i_n j} / I_{i_m j}$$

Proteomics Informatics – Protein quantitation II: software (Week 9)



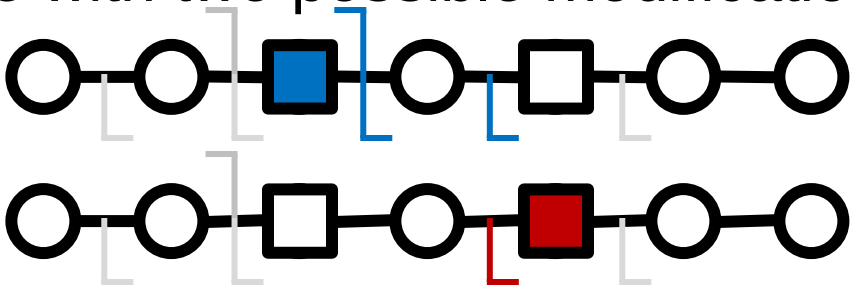
MaxQuant

Skyline



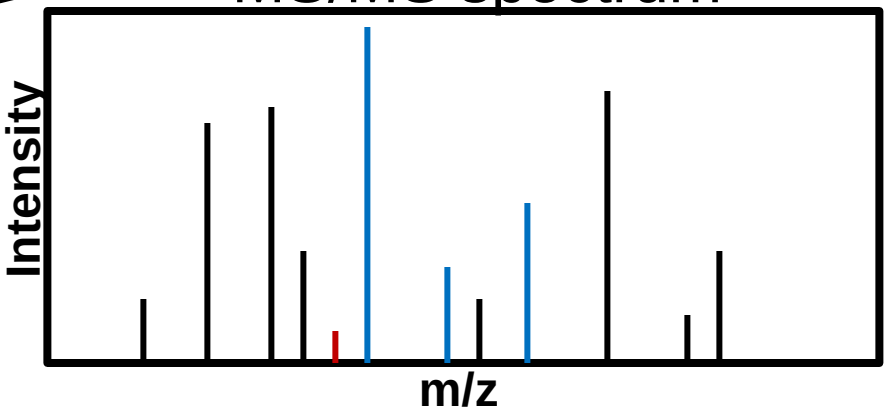
Proteomics Informatics – Protein characterization I: post-translational modifications (Week 10)

Peptide with two possible modification sites



Matching

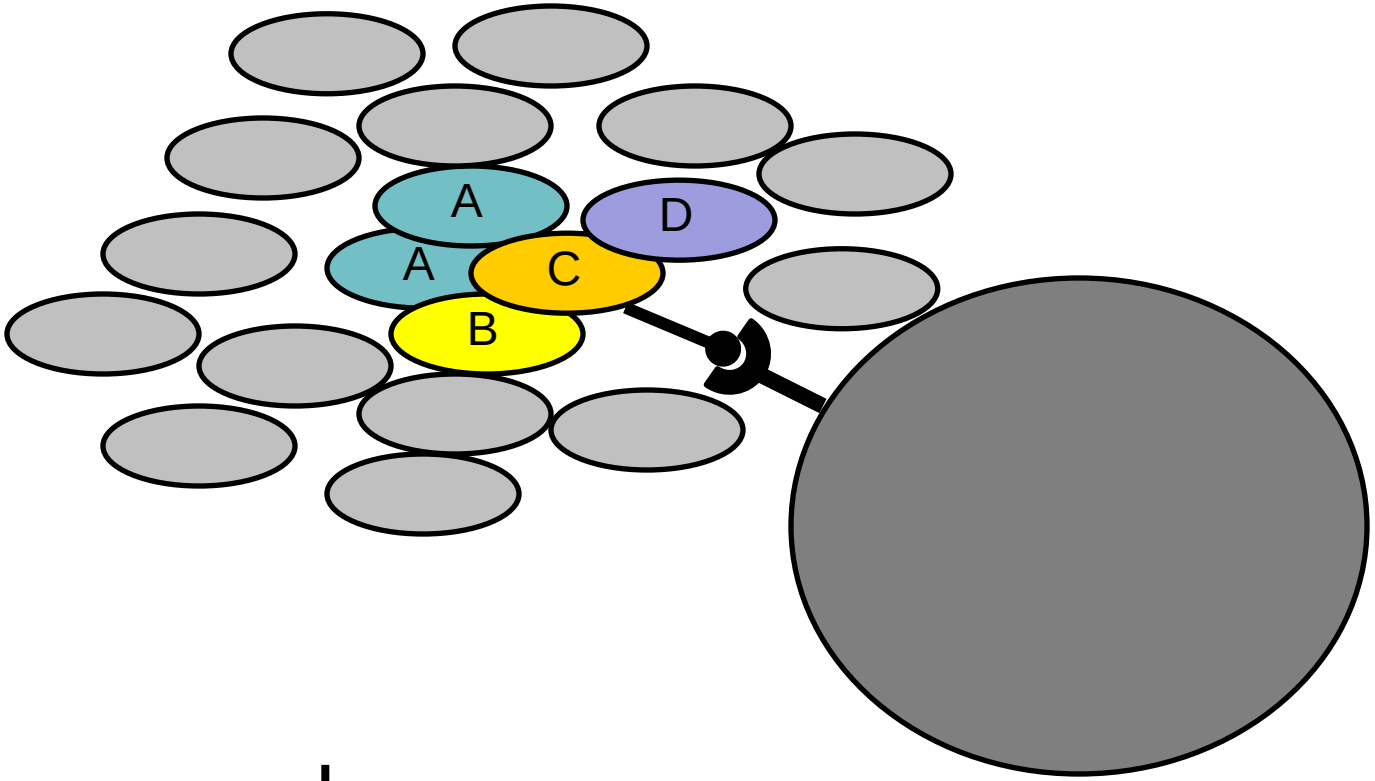
MS/MS spectrum



Which assignment does the data support?

1, 1 or 2, or 1 and 2?

Proteomics Informatics – Protein Characterization II: protein-protein interactions, cross-linking, top-down, non-covalent complexes (Week 11)

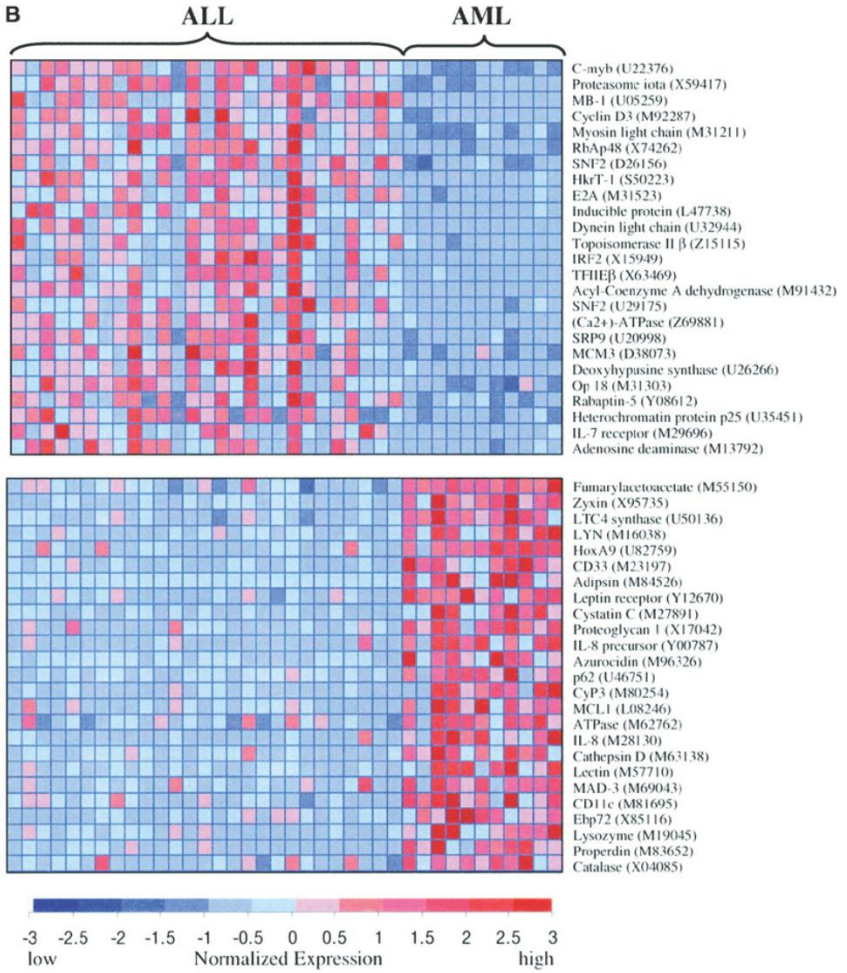


↓

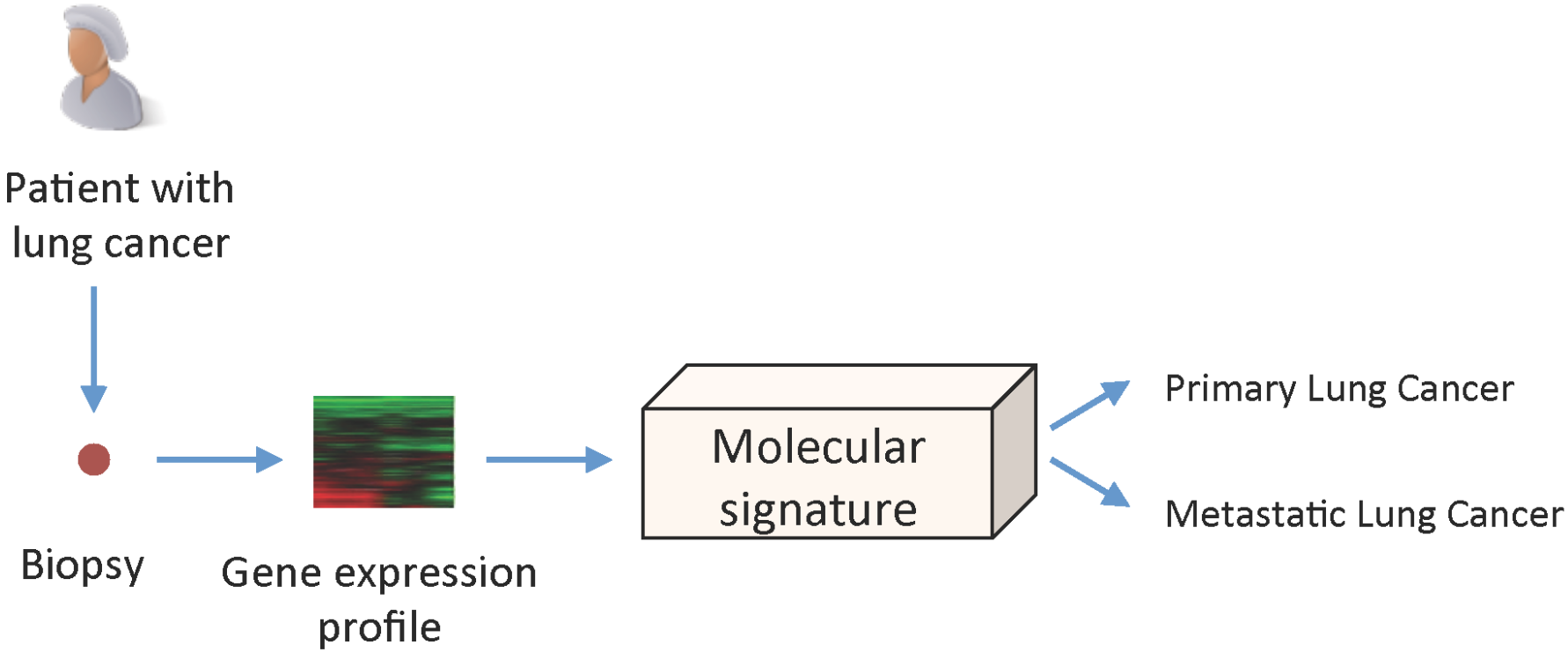
Protein identification

Proteomics Informatics – Molecular Signatures (Week 12)

Molecular signature is a computational or mathematical model that links high-dimensional molecular information to phenotype or other response variable of interest.



Proteomics Informatics – Molecular Signatures (Week 12)



Proteomics Informatics – Presentations of projects (Week 13)

Select a published data set that has been made public and reanalyze it.

Highlighted data sets: http://www.thegpm.org/

10 min presentations

Data sets of the year (2012/12/30)
Technical, Biological and Clinical.

This week we are highlighting the three finest examples of proteomics data made public in 2012. As we have been doing for several years, we are naming the best data in three categories. N.B., these ratings do not take into account the associated publication: only the data itself was considered in these awards. Any of these data sets would be ideal for use as standards in the development of any type of bioinformatics or computational biology algorithms associated with proteomics data.

1. **Technical data:**
   Vaudel M, Burkhart JM, Radau S, Zahedi RP, Martens L and Sickmann A
   Integral Quantification Accuracy Estimation for Reporter Ion-based Quantitative Proteomics (iQuARI). (link)
   Excellent data quality, laboratory technique and an unusual take on quantitation all contributed to selecting this data set.
2. **Biological data:**
   Bischof S, Baerenfaller K, Wildhaber T, Troesch R, Vidi PA, Roschitzki B, Hirsch-Hoffmann M, Hennig L, Kessler F, Gruissem W, and Baginsky S
   Plastid proteome assembly without Toc159: photosynthetic protein import and accumulation of N-acetylated plastid precursor proteins. (link)
   A truly outstanding data set attempting to solve a difficult biological problem. The exploration of how chloroplasts are generated and function within plants is a key biological problem and these results were an example of some of the best practices to address the associated protein trafficing issues.
3. **Clinical data:**
   Steiling K, Kadar AY, Bergerat A, Flanigon J, Sridhar S, Shah V, Ahmad QR, Brody JS, Lenburg ME, Steffen M, and Spira A
   Comparison of proteomic and transcriptomic profiles in the bronchial airway epithelium of current and never smokers. (link)
   Working with clinical tissues is still a challenge for many groups in proteomics, but this study demonstrated that consistently excellent data can be obtained even when working with relatively large, heterogenous populations.

Proteomics Informatics (BMSC-GA 4437)

Instructor

David Fenyö

Contact information

David@FenyoLab.org

http://fenyolab.org/presentations/Proteomics_Informatics_2013/