Proteomics Informatics (BMSC-GA 4437)

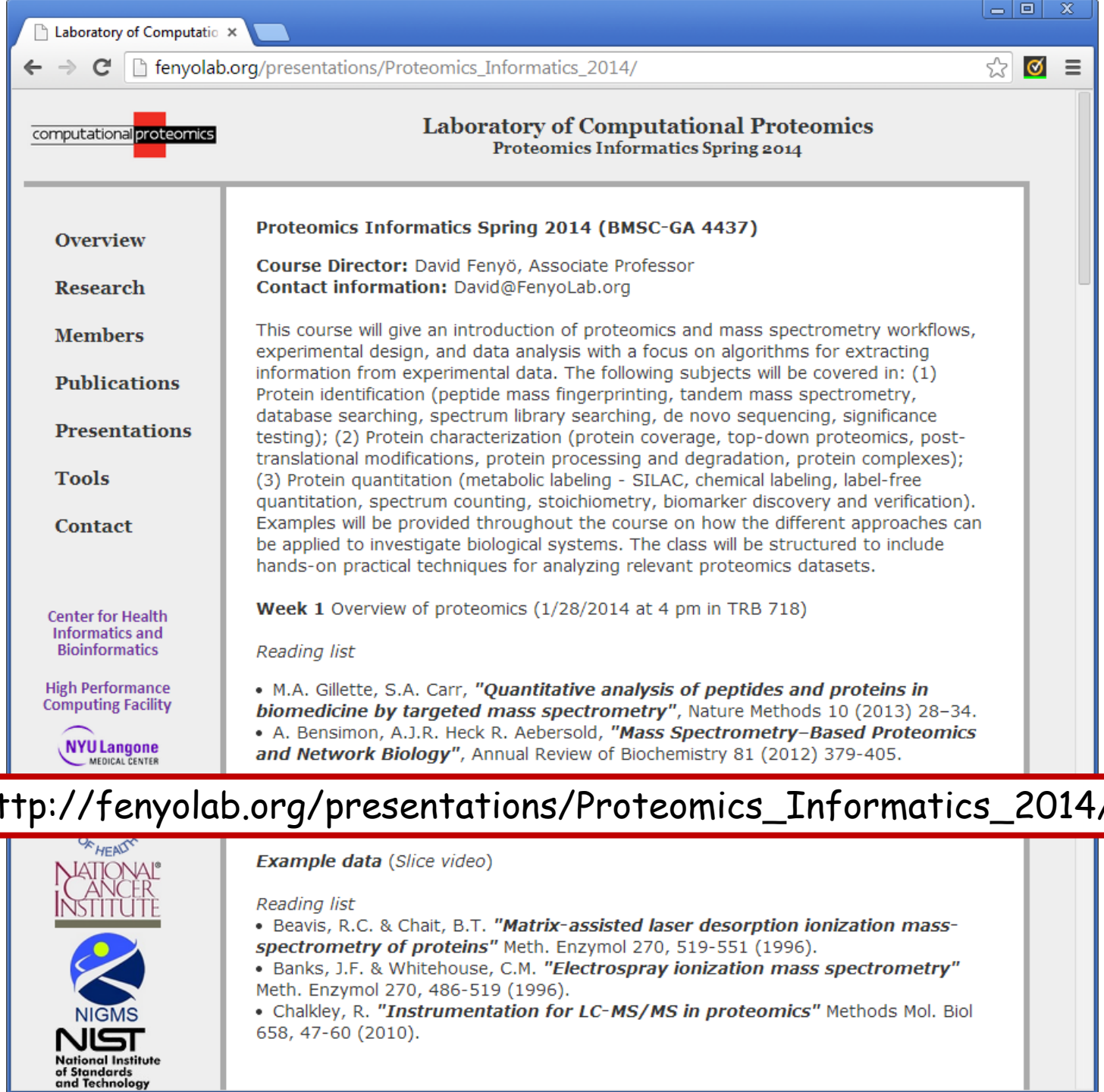
Course Director

David Fenyö

Contact information

David@FenyoLab.org

http://fenyolab.org/presentations/Proteomics_Informatics_2014/


Laboratory of Computatio
fenyolab.org/presentations/Proteomics_Informatics_2014/
computational proteomics
Laboratory of Computational Proteomics
Proteomics Informatics Spring 2014
Overview
Research
Members
Publications
Presentations
Tools
Contact
Center for Health Informatics and Bioinformatics
High Performance Computing Facility
NYU Langone MEDICAL CENTER
NATIONAL CANCER INSTITUTE
NIGMS
NIST National Institute of Standards and Technology
Proteomics Informatics Spring 2014 (BMSC-GA 4437)
Course Director: David Fenyö, Associate Professor
Contact information: David@FenyoLab.org
This course will give an introduction of proteomics and mass spectrometry workflows, experimental design, and data analysis with a focus on algorithms for extracting information from experimental data. The following subjects will be covered in: (1) Protein identification (peptide mass fingerprinting, tandem mass spectrometry, database searching, spectrum library searching, de novo sequencing, significance testing); (2) Protein characterization (protein coverage, top-down proteomics, post-translational modifications, protein processing and degradation, protein complexes); (3) Protein quantitation (metabolic labeling - SILAC, chemical labeling, label-free quantitation, spectrum counting, stoichiometry, biomarker discovery and verification). Examples will be provided throughout the course on how the different approaches can be applied to investigate biological systems. The class will be structured to include hands-on practical techniques for analyzing relevant proteomics datasets.
Week 1 Overview of proteomics (1/28/2014 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.
http://fenyolab.org/presentations/Proteomics_Informatics_2014/
Example data (Slice video)
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).

Proteomics Informatics – Learning Objectives

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

Proteomics Informatics – Syllabus

Week 1 Overview of proteomics (1/28/2014 at 4 pm in TRB 718)

Week 2 Overview of mass spectrometry (2/4/2014 at 4 pm in TRB 718)

Week 3 Analysis of mass spectra: signal processing, peak finding, and isotope clusters (2/11/2014 at 4 pm in TRB 119)

Week 4 Protein identification I: searching protein sequence collections and significance testing (2/18/2014 at 4 pm in TRB 718)

Week 5 Protein identification II: de novo sequencing (2/25/2014 at 4 pm in TRB 718)

Week 6 Databases, data repositories and standardization (3/4/2014 at 4 pm in TRB 718)

Week 7 Proteogenomics (3/11/2014 at 4 pm in TRB 718)

Week 8 Protein quantitation I: Overview (3/18/2014 at 4 pm in TRB 718)

Week 9 Protein quantitation II: Targeted (3/25/2014 at 4 pm in TRB 718)

Week 10 Protein characterization I: post-translational modifications (4/1/2014 at 4 pm in TRB 718)

Week 11 Protein characterization II: Protein interactions (4/10/2014 at 4 pm in TRB 718)

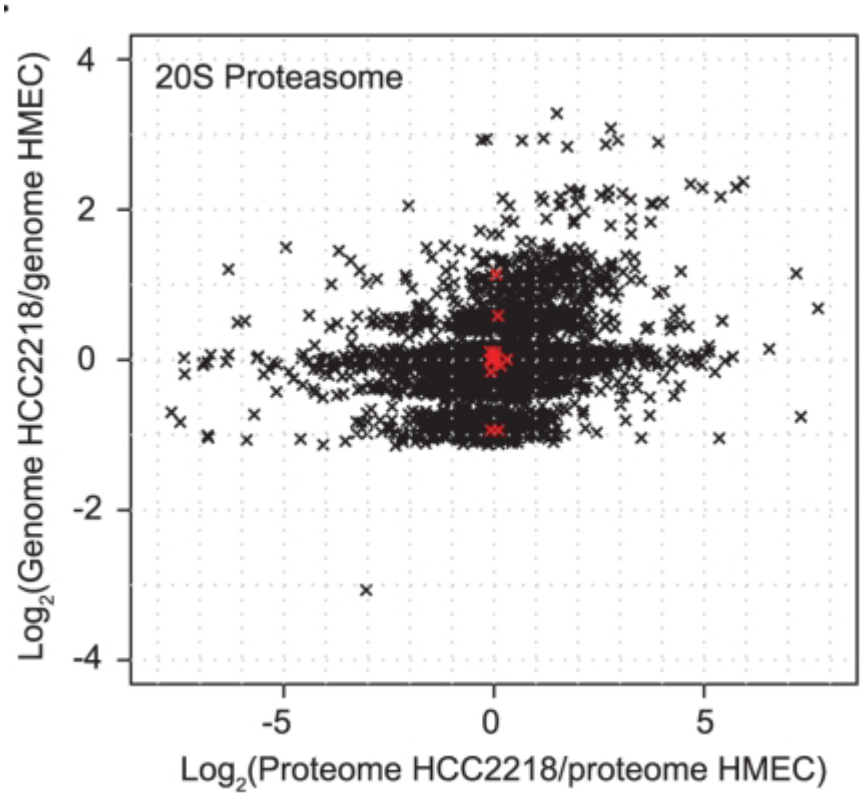
Week 12 Molecular Signatures (4/17/2014 at 4 pm in TRB 718)

Week 13 Presentations of projects (4/22/2014 at 4 pm in TRB 718)

Proteomics Informatics – Overview of Proteomics (Week 1)

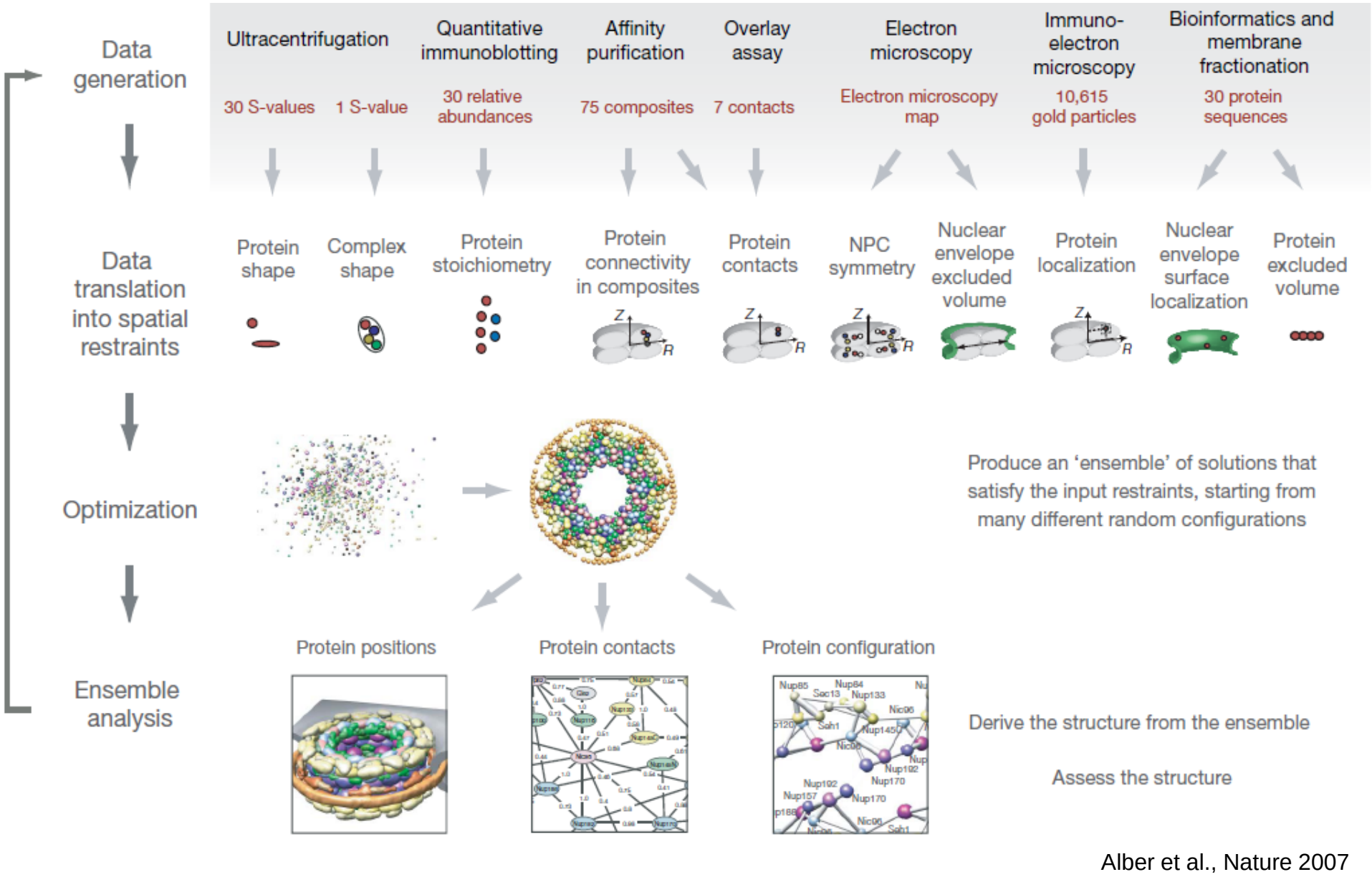
- Why proteomics?
- Bioinformatics
- Overview of the course

Motivating Example: Protein Regulation



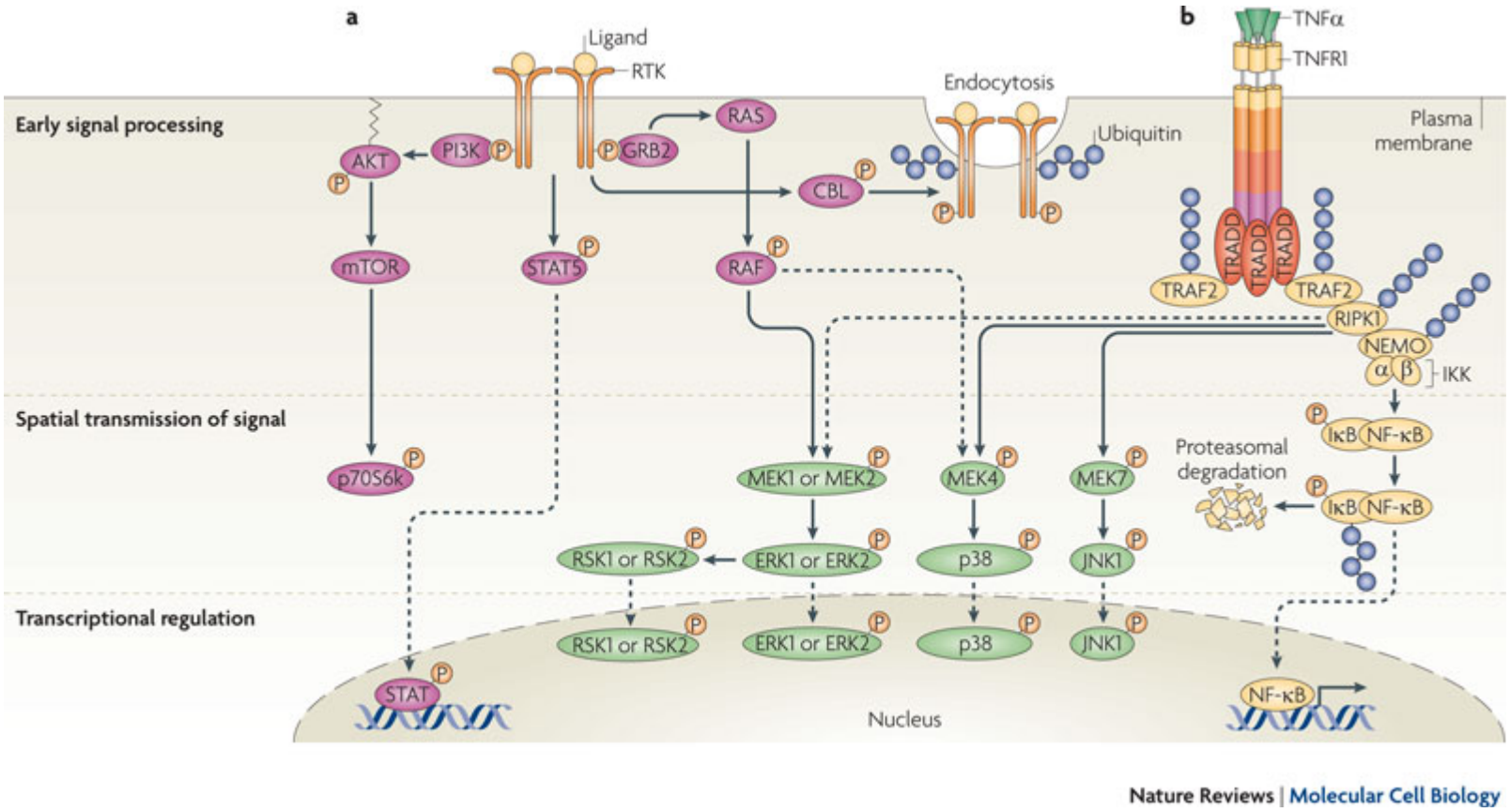
Geiger et al., “Proteomic changes resulting from gene copy number variations in cancer cells”, PLoS Genet. 2010 Sep 2;6(9). pii: e1001090.

Motivating Example: Protein Complexes



Alber et al., Nature 2007

Motivating Example: Signaling

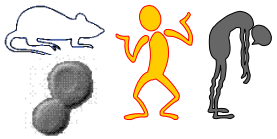


Nature Reviews | Molecular Cell Biology

Choudhary & Mann, Nature Reviews Molecular Cell Biology 2010

Bioinformatics

Biological System



↓ Experimental Design

Samples

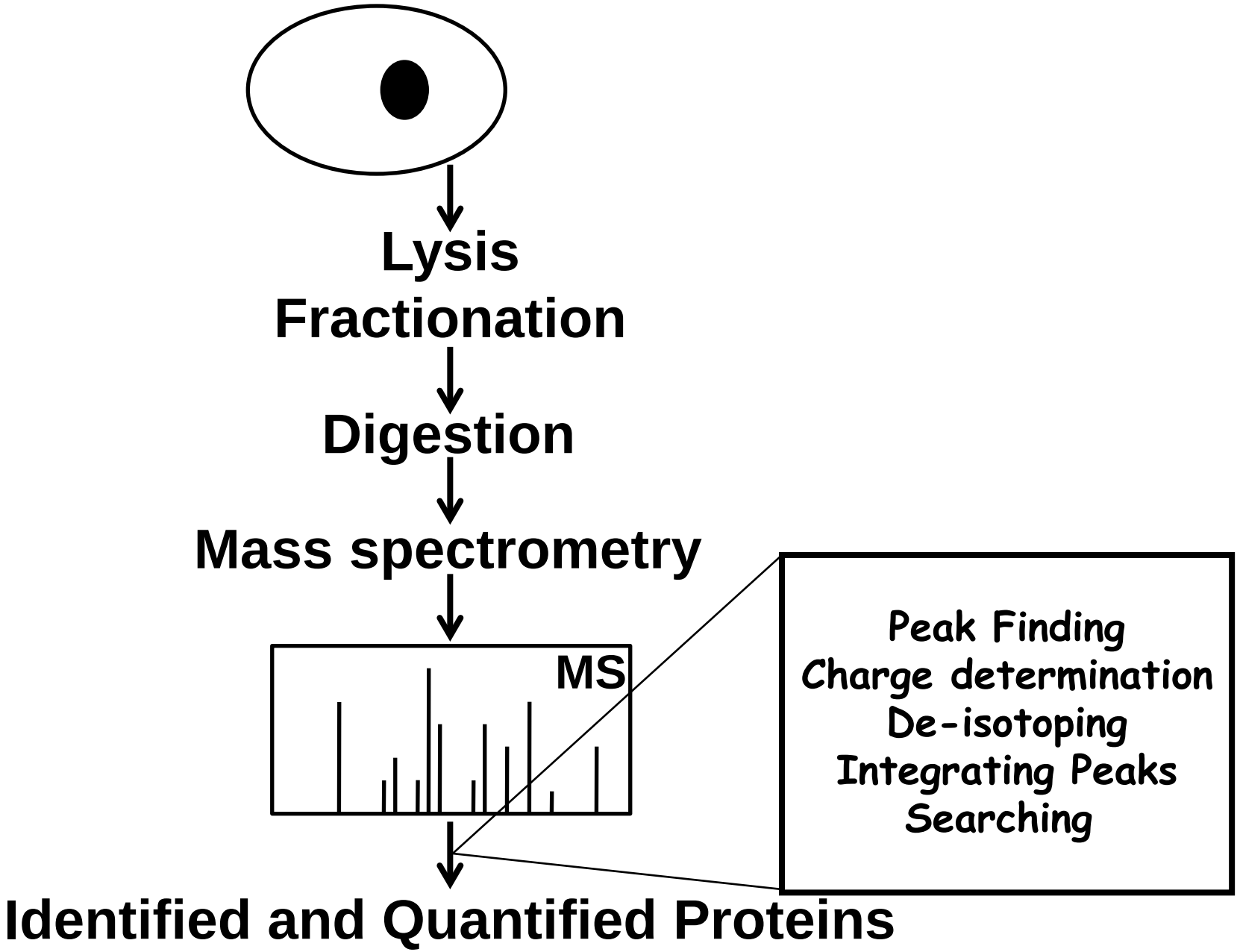
↓ Measurements

Raw Data

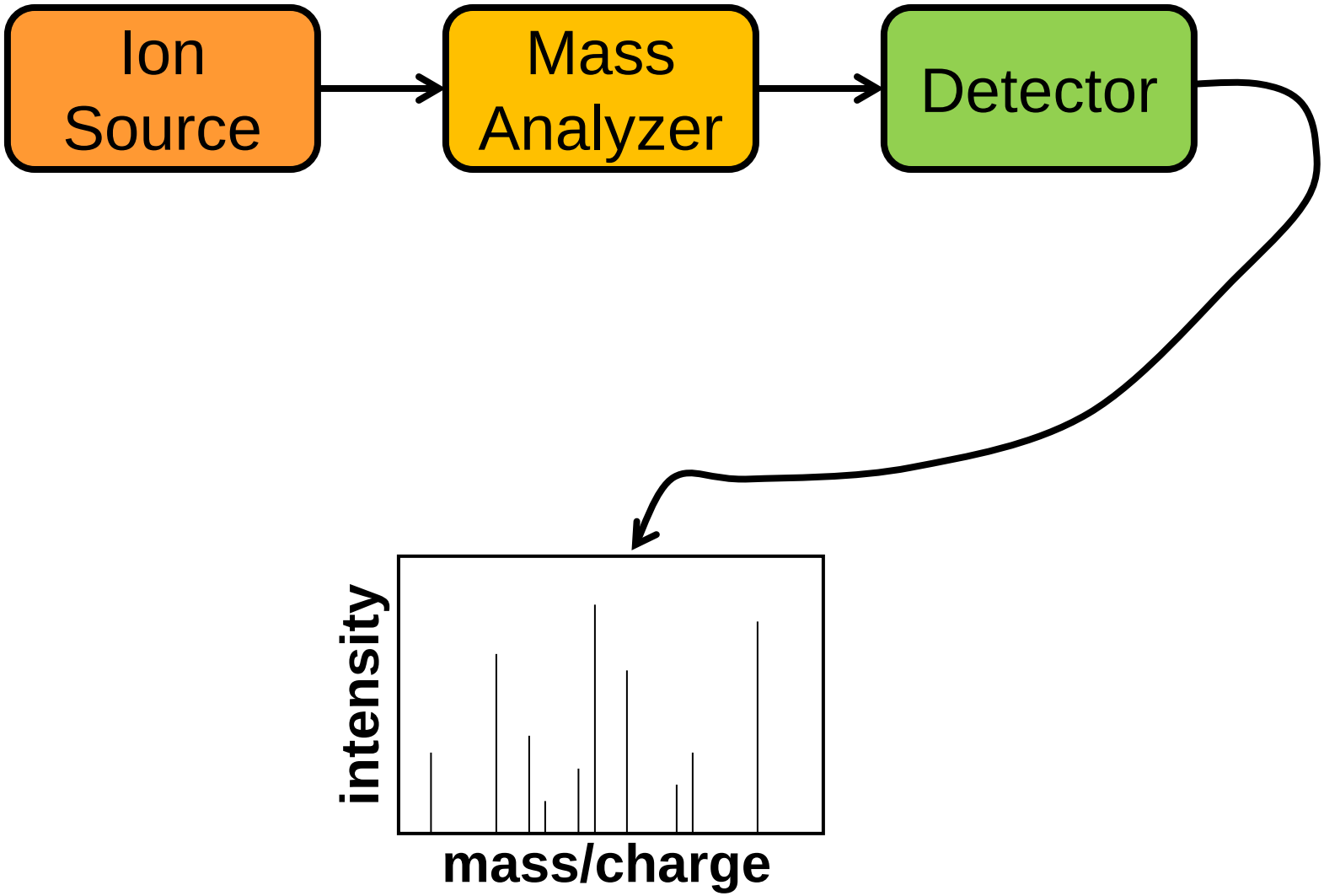
↓ Data Analysis

Information

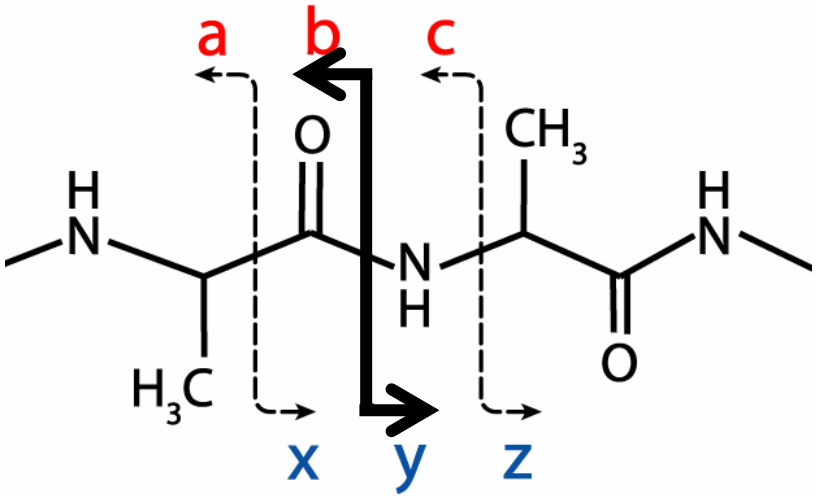
Mass Spectrometry Based Proteomics

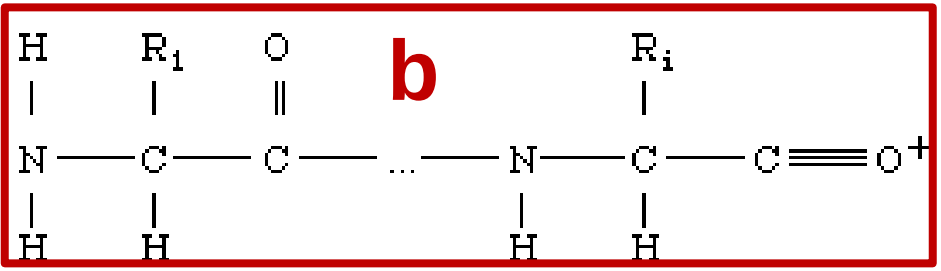


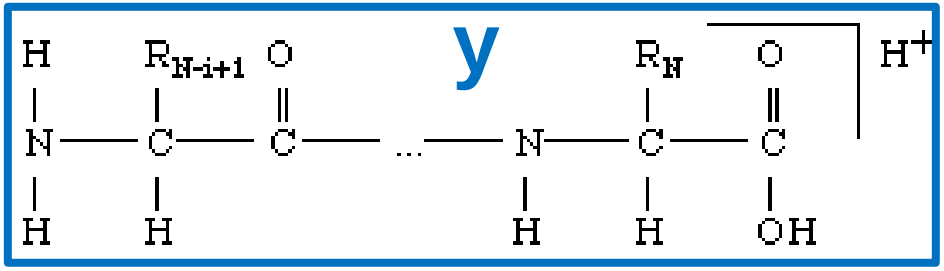
Proteomics Informatics – Overview of Mass spectrometry (Week 2)



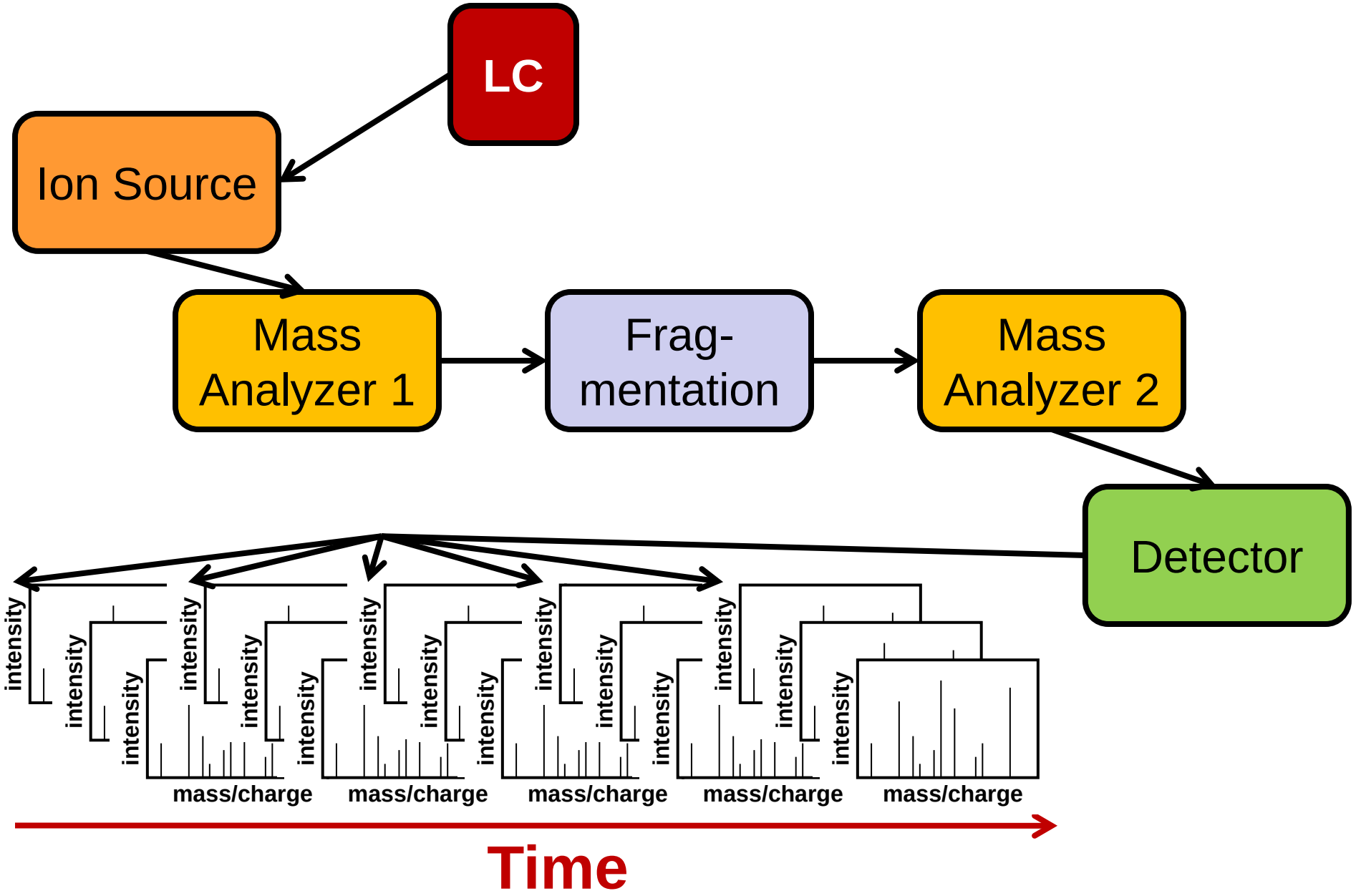
Proteomics Informatics – Overview of Mass spectrometry (Week 2)



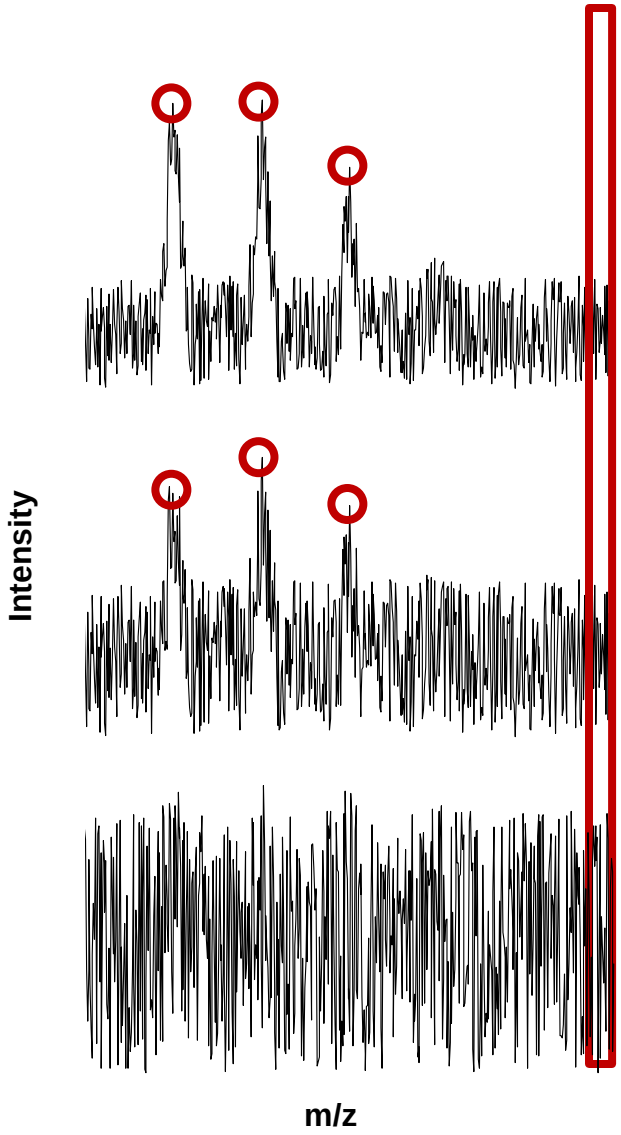




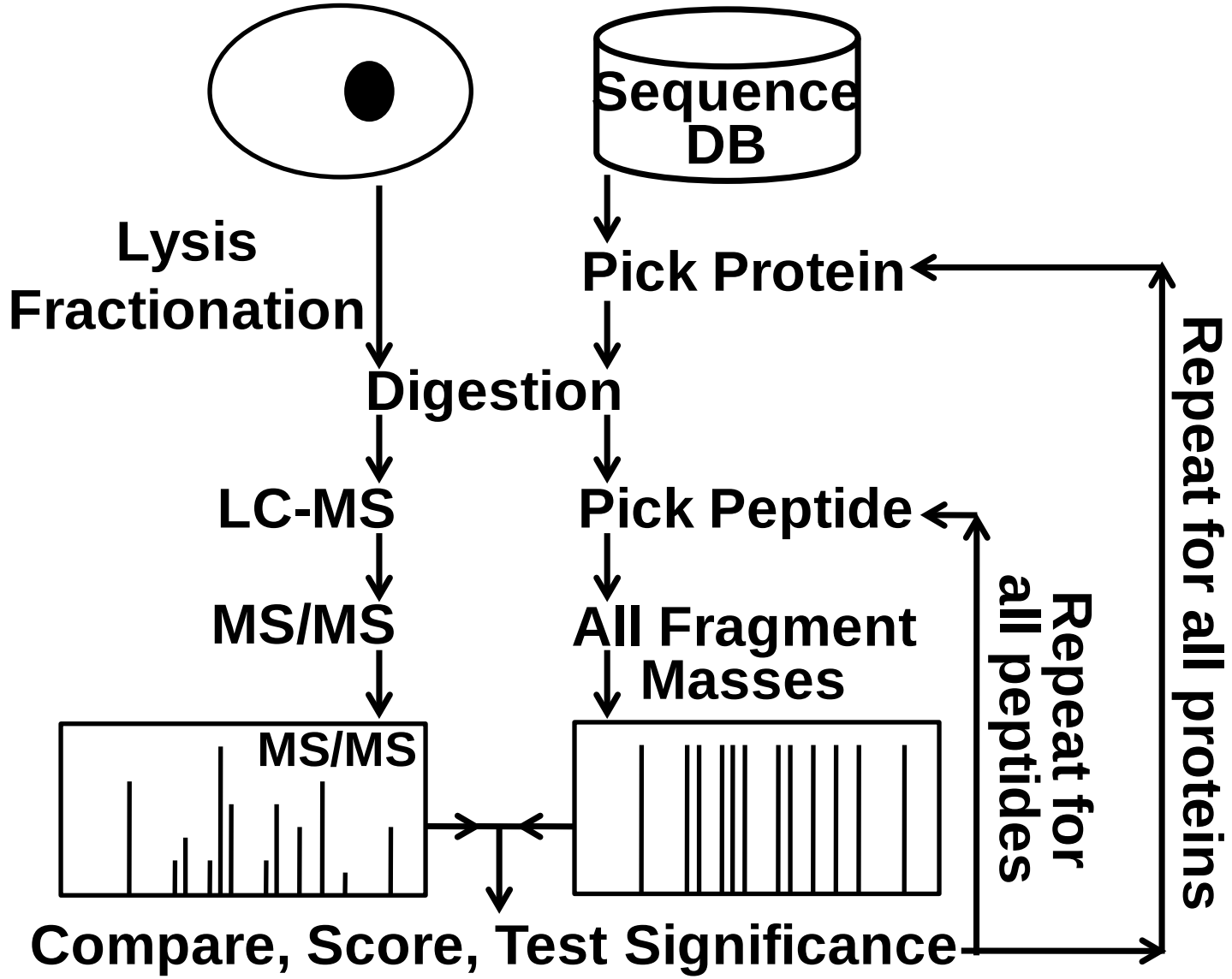
Proteomics Informatics – Overview of Mass spectrometry (Week 2)



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)



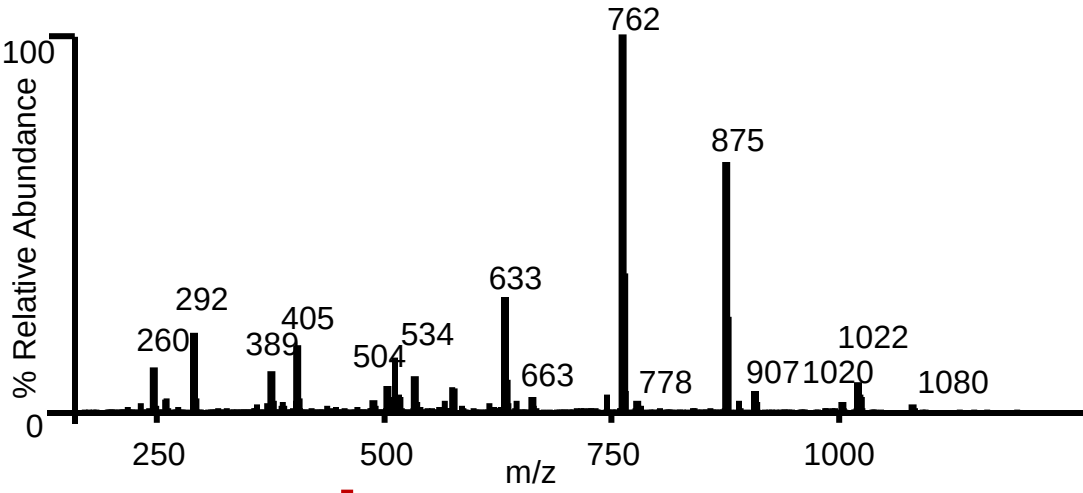
Proteomics Informatics – Protein identification I: searching protein sequence collections and significance testing (Week 4)

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

Proteomics Informatics – Protein identification II: de novo sequencing (Week 5)

Amino acid masses

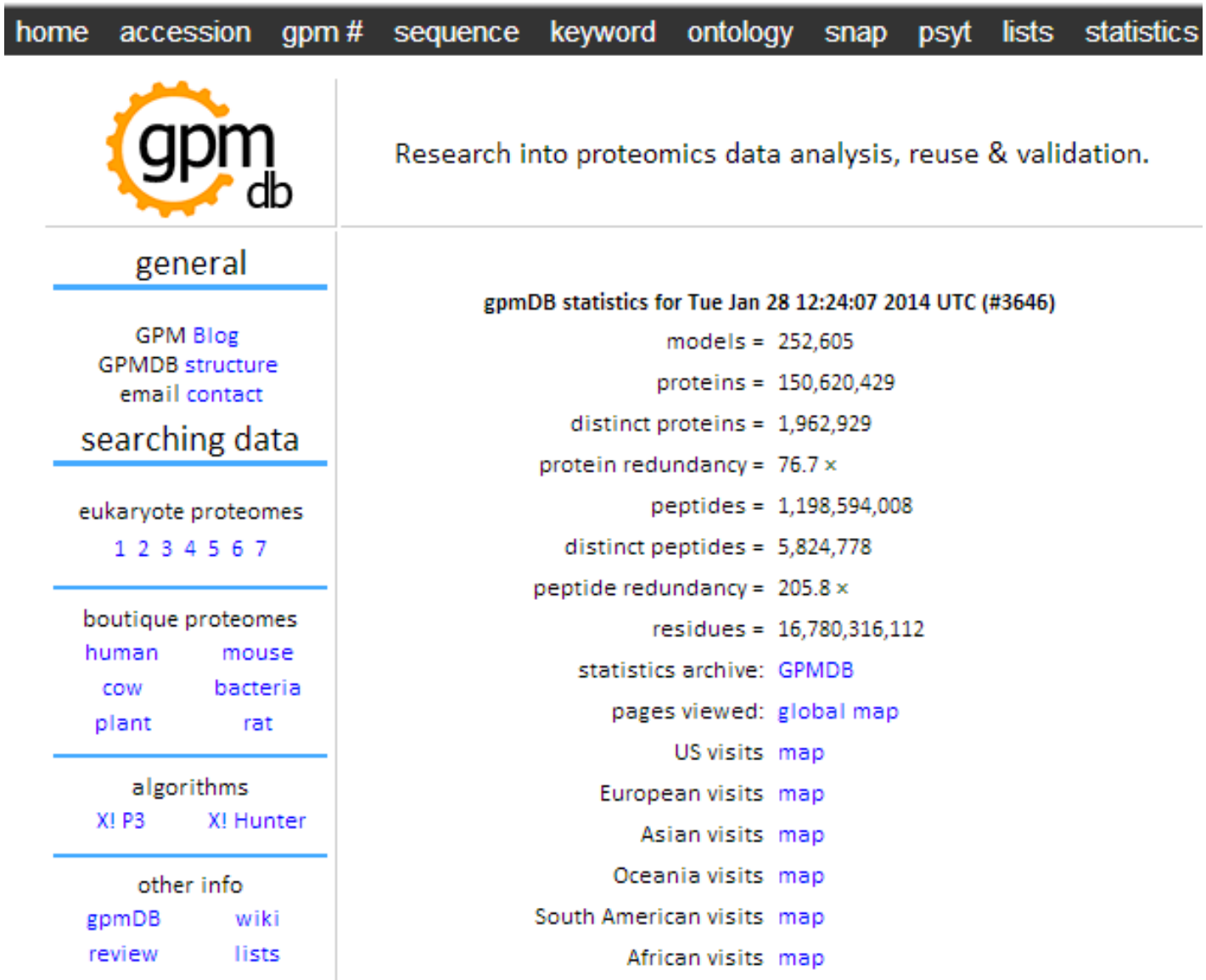
1-letter code	3-letter code	Chemical formula	Monoisotopic	Average
A	Ala	C_3H_5ON	71.0371	71.0788
R	Arg	$C_6H_{12}ON_4$	156.101	156.188
N	Asn	$C_4H_6O_2N_2$	114.043	114.104
D	Asp	$C_4H_5O_3N$	115.027	115.089
C	Cys	C_3H_5ONS	103.009	103.139
E	Glu	$C_5H_7O_3N$	129.043	129.116
Q	Gln	$C_5H_8O_2N_2$	128.059	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



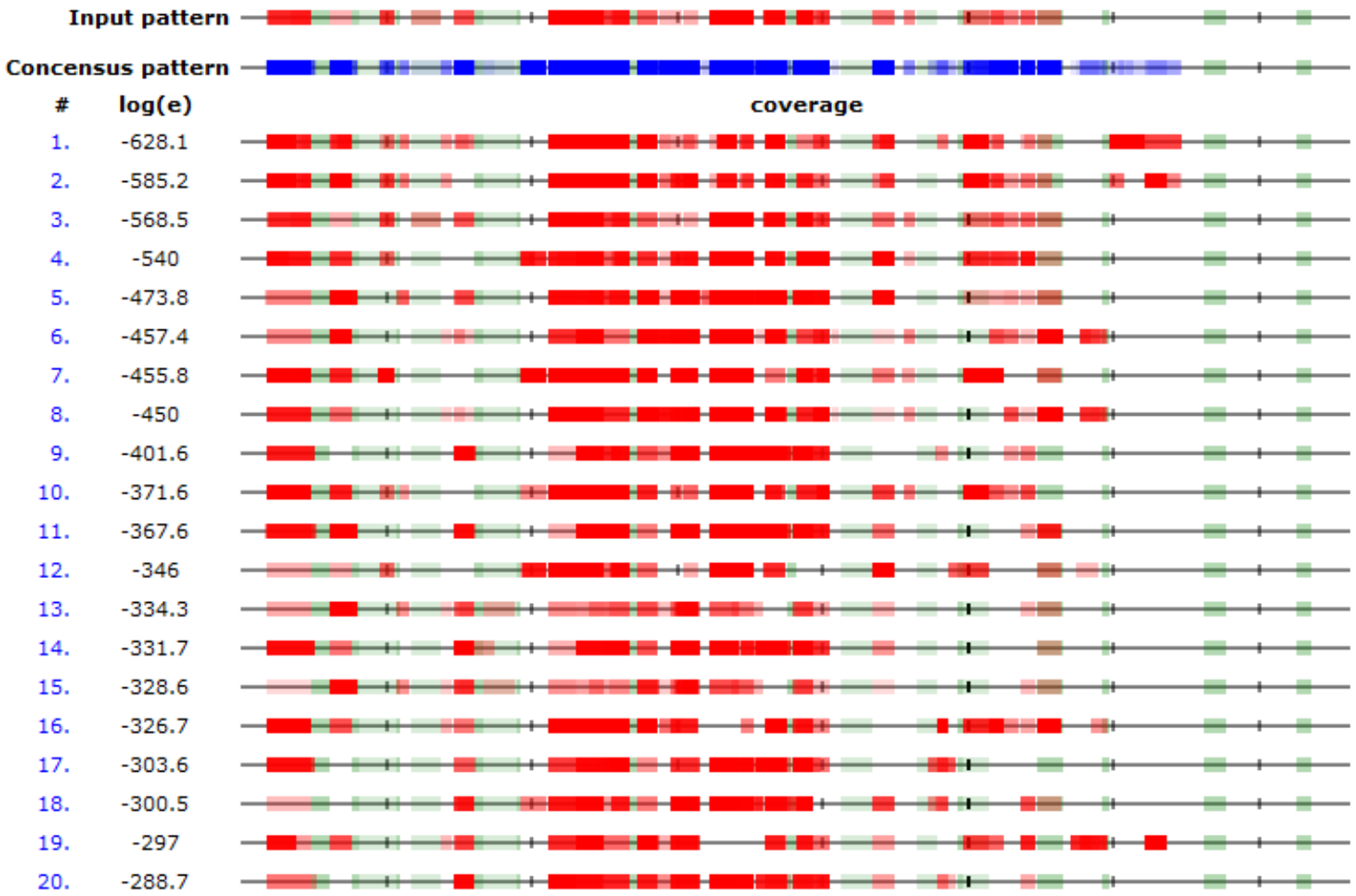
Mass Differences

Sequences consistent with spectrum

Proteomics Informatics – Databases, data repositories and standardization (Week 6)

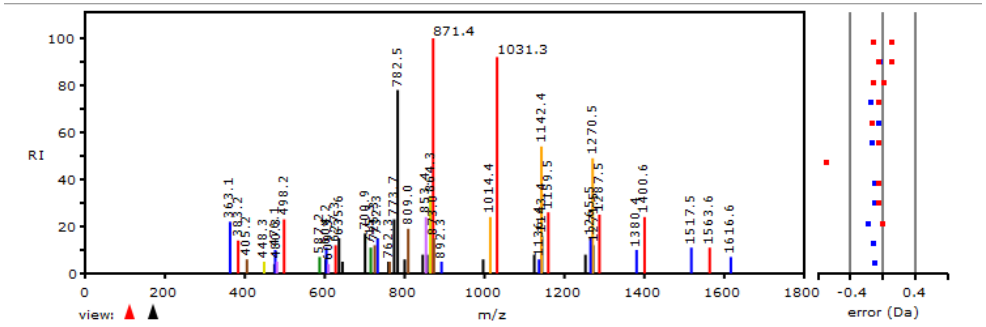


Proteomics Informatics – Databases, data repositories and standardization (Week 6)



Most proteins show very reproducible peptide patterns

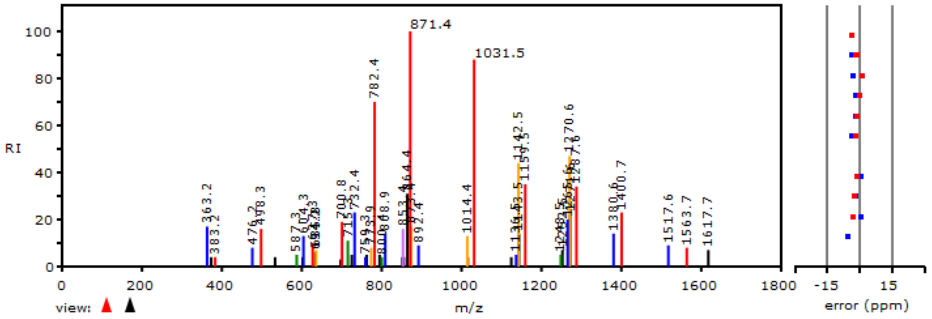
Proteomics Informatics – Databases, data repositories and standardization (Week 6)



1. **cos(θ) = 0.98**, z = 2, log(e) = -14.8, m+h = 1762.8218 (P)

A Q Y L Q Q C P F E D H V K

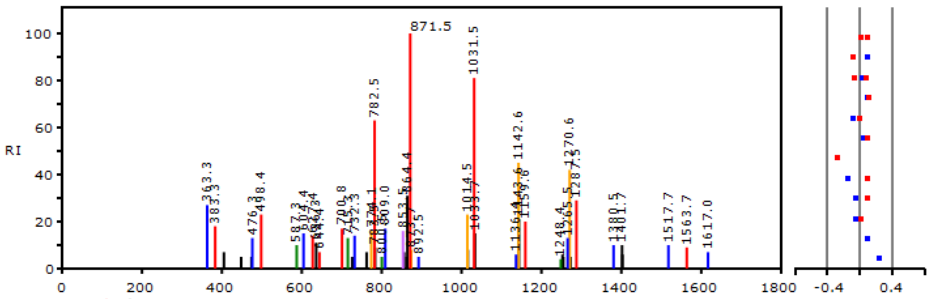
Query Spectrum



2. **cos(θ) = 0.96**, z = 2, log(e) = -13.5, m+h = 1762.8216 (P)

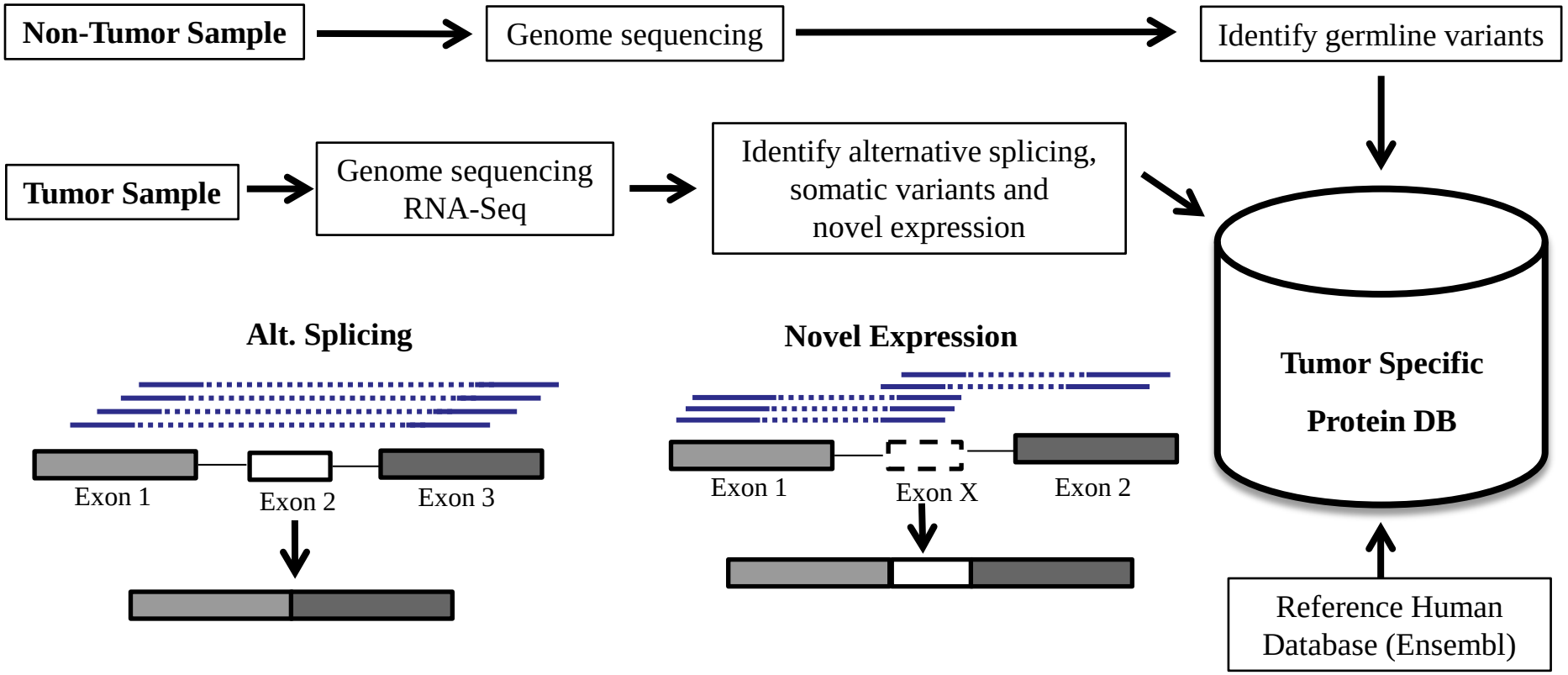
A Q Y L Q Q C P F E D H V K

Best match In GPMDB



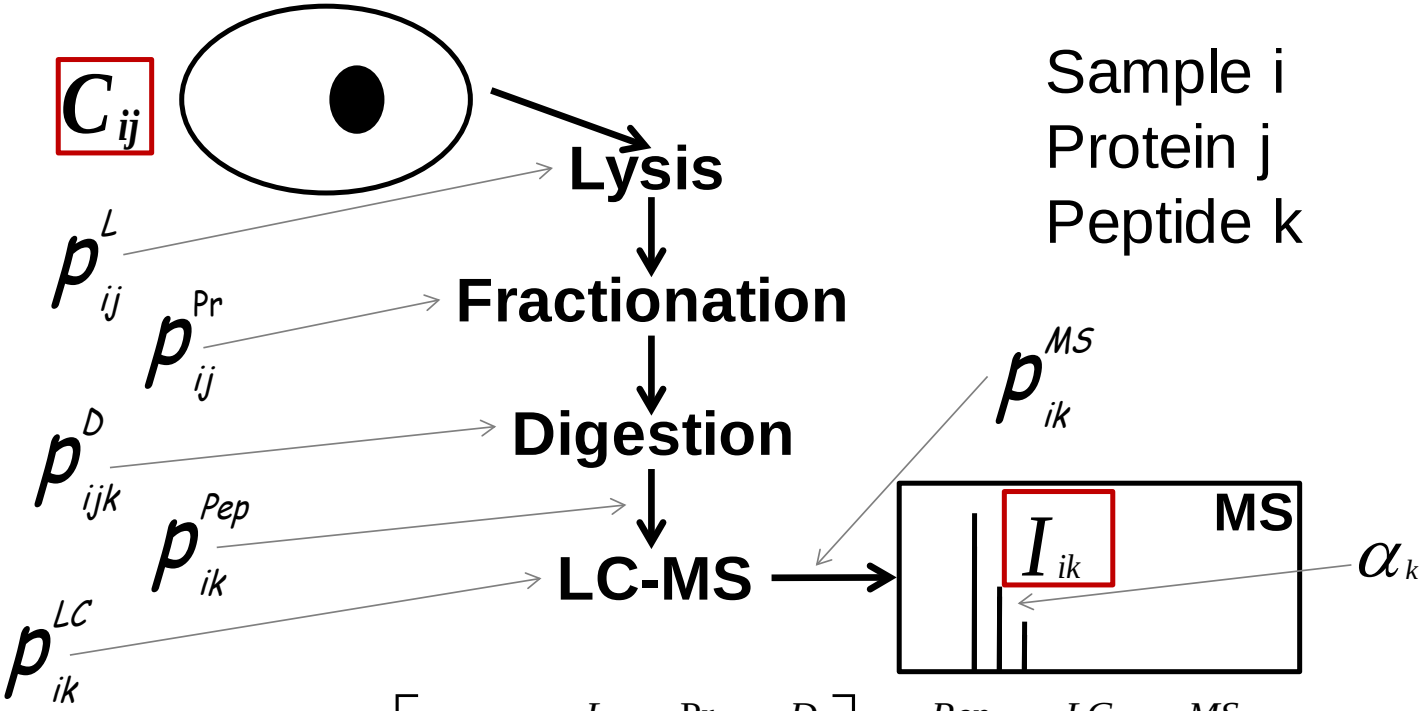
Second best match In GPMDB

Proteomics Informatics - Proteogenomics (Week 7)



Kelly Ruggles

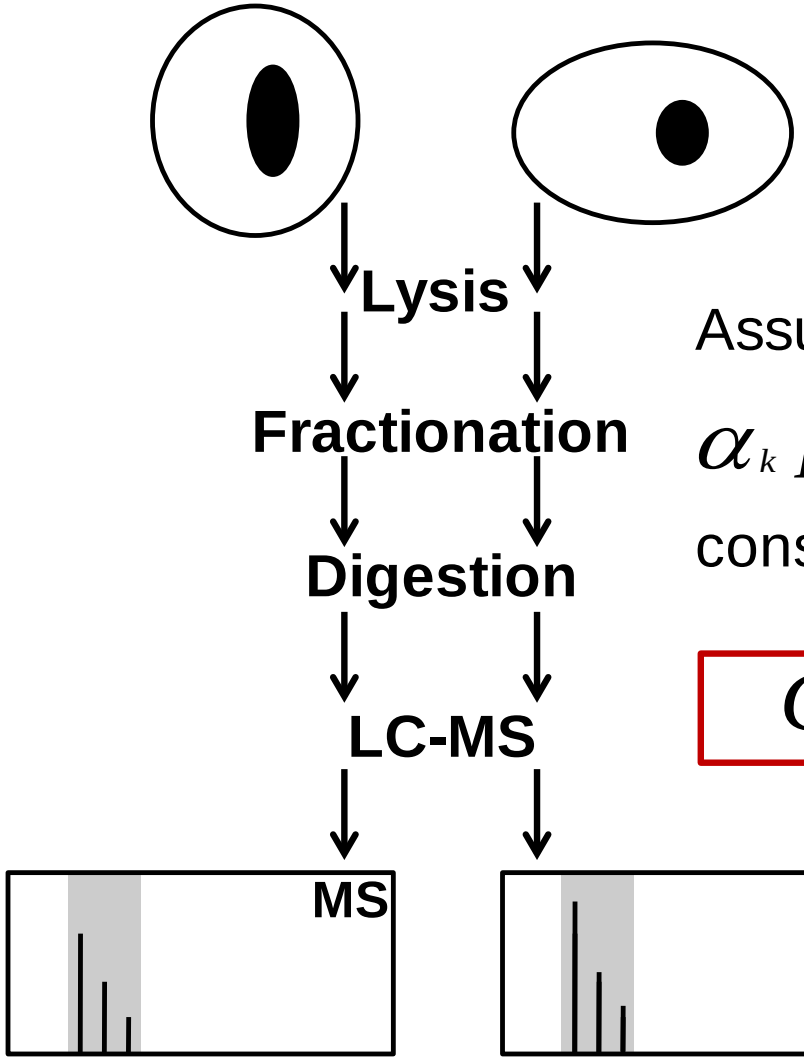
Proteomics Informatics – Protein quantitation I: Overview (Week 8)



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

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

Proteomics Informatics – Protein quantitation I: Overview (Week 8)



Sample i
Protein j
Peptide k

Assumption:

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

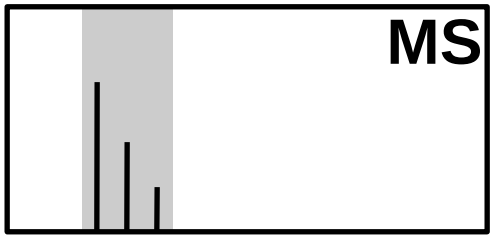
constant for all samples

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

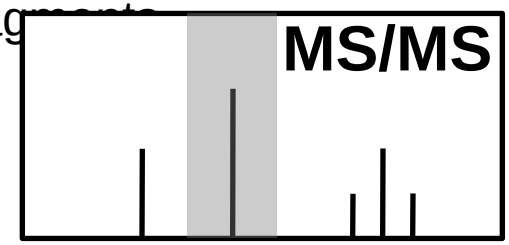
Proteomics Informatics – Protein quantitation II: Targeted (Week 9)

Shotgun proteomics ← **LC-MS** → **Targeted MS**

1. Records M/Z



2. Selects peptides based on abundance and frag



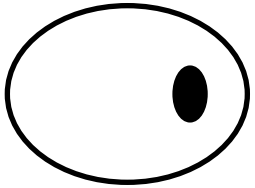
3. Protein database search for peptide identification

Data Dependent Acquisition (DDA)

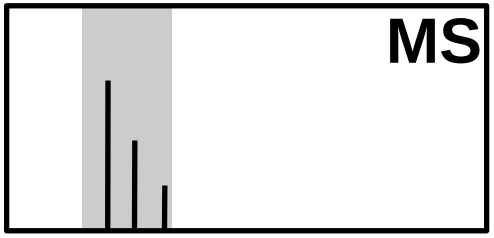
Digestion

Fractionation

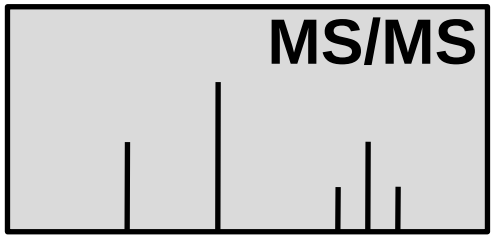
Lysis



1. Select precursor ion



2. Precursor fragmentation

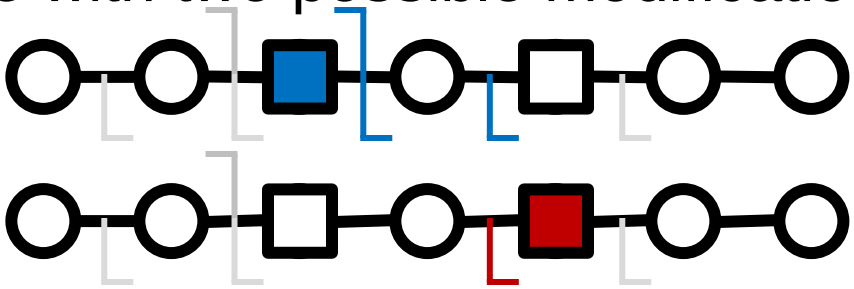


3. Use Precursor-Fragment pairs for identification

Uses predefined set of peptides

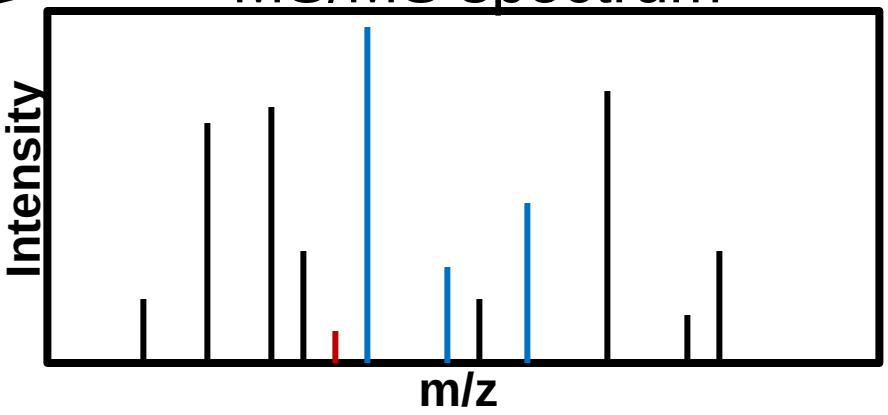
Proteomics Informatics – Protein characterization I: post-translational modifications (Week 10)

Peptide with two possible modification sites



Matching

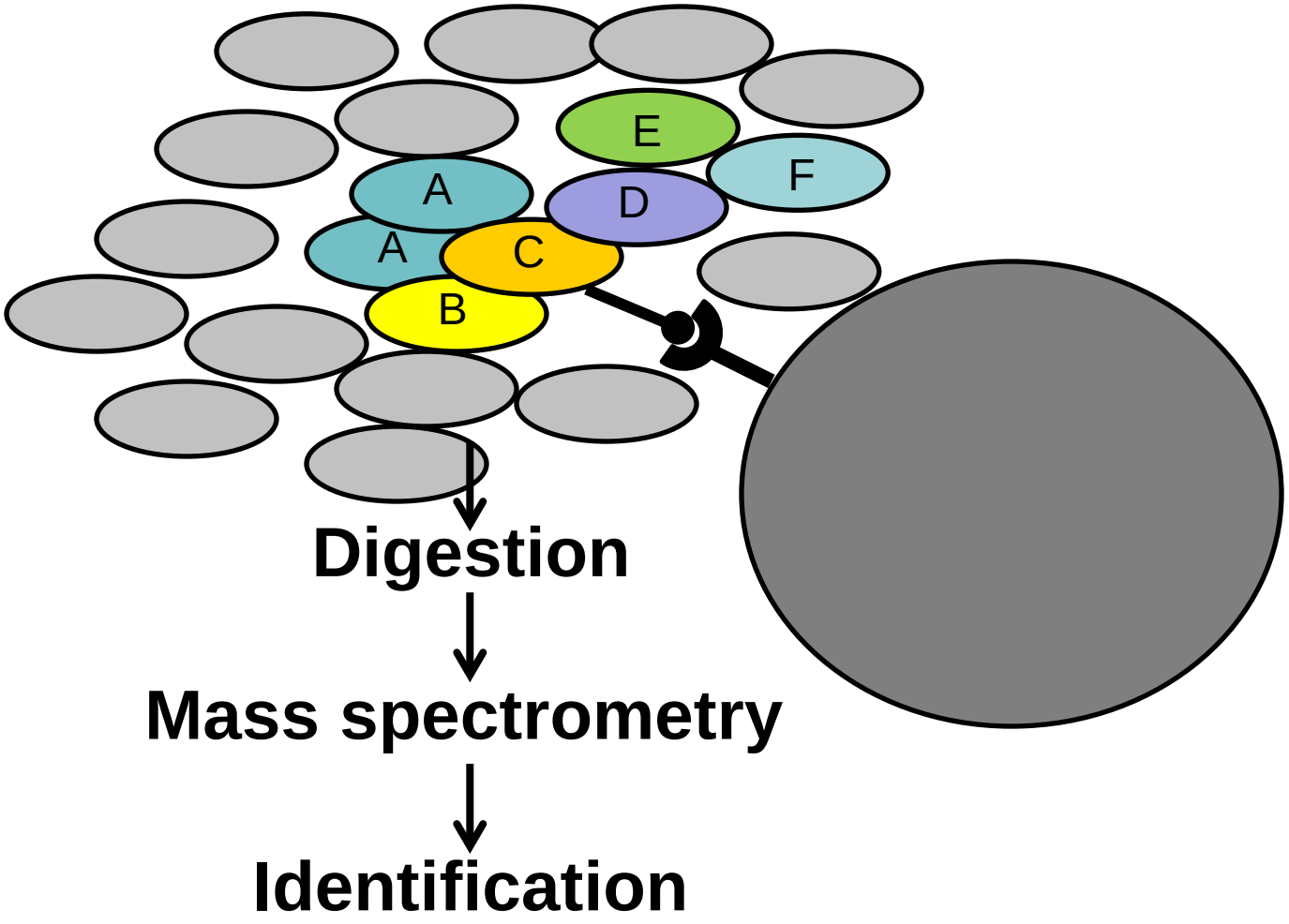
MS/MS spectrum



Which assignment does the data support?

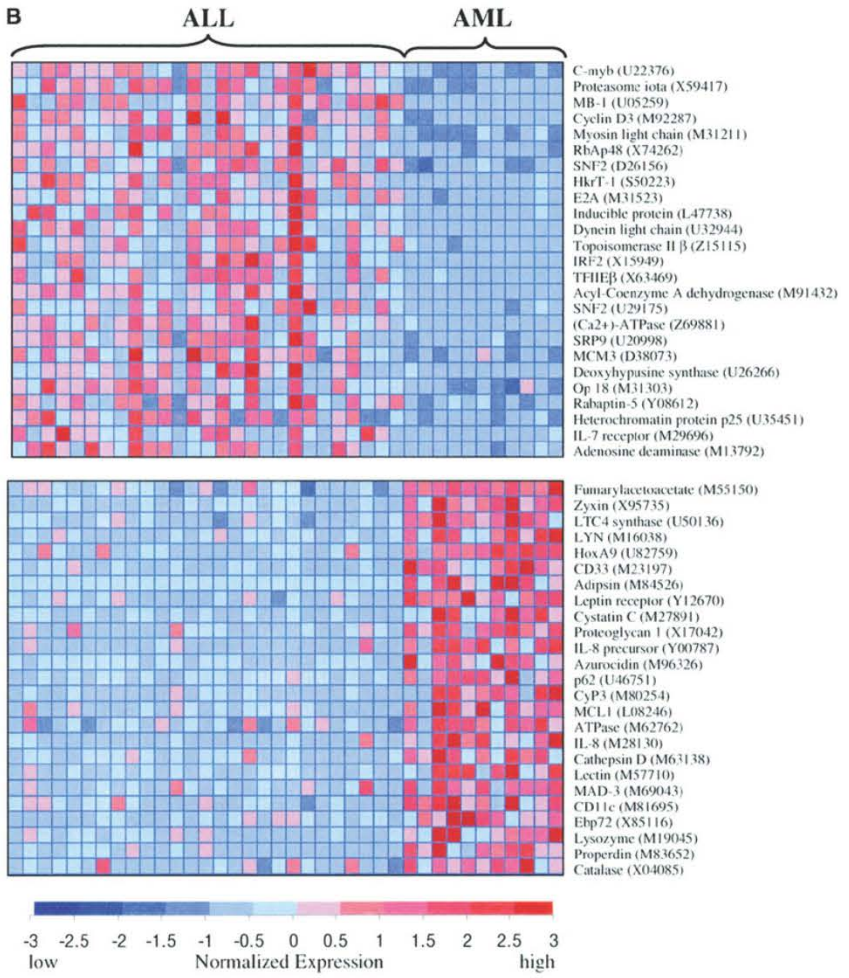
1, 1 or 2, or 1 and 2?

Proteomics Informatics – Protein Characterization II: protein interactions (Week 11)

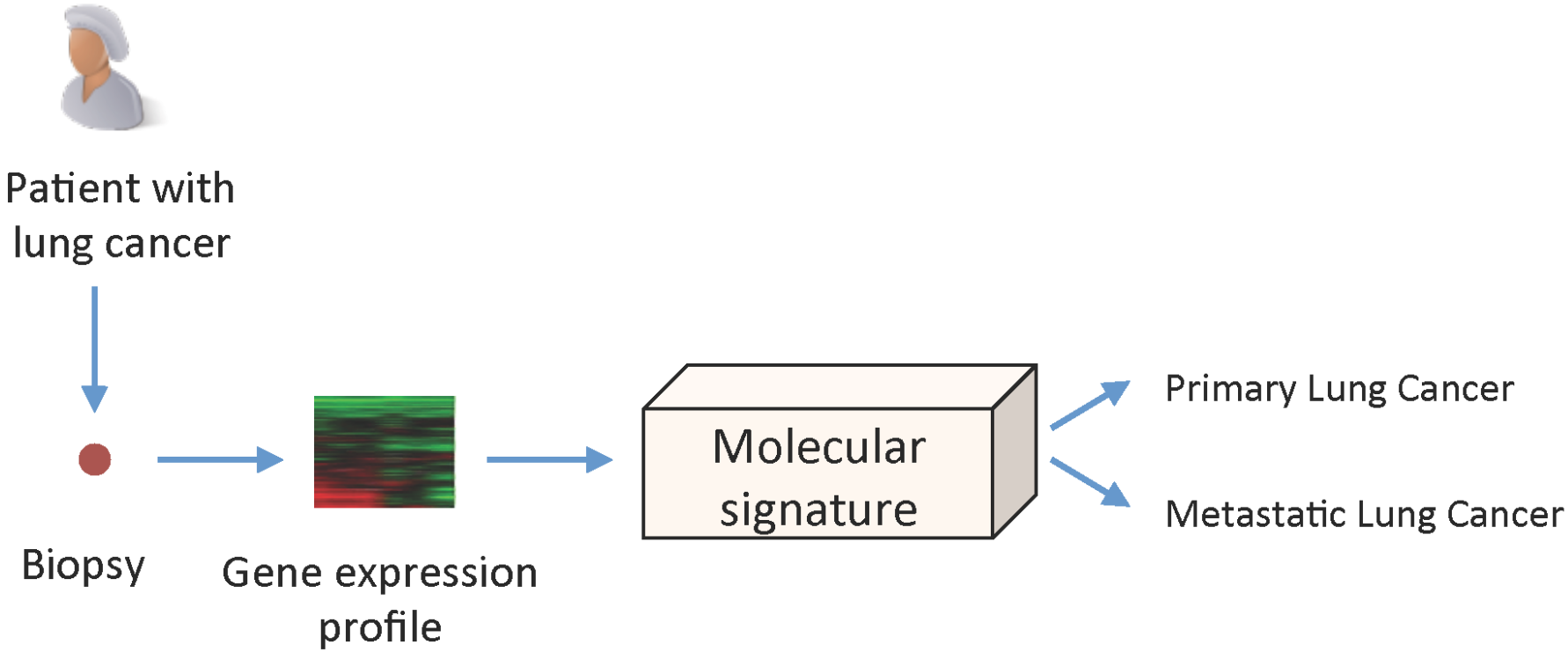


Proteomics Informatics – Molecular Signatures (Week 12)

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

The Global Proteome Machine

Proteomics data analysis, reuse and validation for biological and biomedical research.

GPM Blog

Data set of the week: (2014/1/26)
Proteomic analysis of purified protein derivative of *Mycobacterium tuberculosis*.
Overall rating: ★★ very good data (general interest)

This data set consisted of 1 result, a single injection LC/MS/MS experiment. The data file was made available through ProteomeXchange, PXD000377. It has been published by Prasad TS, Verma R, Kumar S, Nirujogi RS, Sathe GJ, Madugundu AK, Sharma J, Puttamallesh VN, Ganjiwale A, Myneedu VP, Chatterjee A, Pandey A, Harsha H, and Narayana J, Clin Proteomics. 2013 Jul 19;10(1):8 (PubMed).

While many published 'omics studies focus on the heroic collection of large volumes of data, this study is more of a haiku: a quiet reflection on an important clinical material. By limiting the study to simply looking at the real composition of "Purified Protein Derivative" (the antigenic material used for the tuberulosus skin test), the authors clearly demonstrate both the power of the now-routine techniques employed and beg the question of why this type of analysis is not available for every batch of this product used clinically.

Data set of the week: (2014/1/19)
Comparative Proteome Analysis Revealing an 11-Protein Signature for Aggressive Triple-Negative Breast Cancer.
Overall rating: ★★★★ excellent data (leading the field)

This data set consisted of 126 results, each one a 3 hour gradient LC/MS/MS experiment from laser microdisected samples. The data files were made available through PeptideAtlas, PASS00260. It has been published by Liu NQ, Stingl C, Look MP, Smid M, Braakman RB, De Marchi T, Sieuwerts AM, Span PN, Sweep FC, Linderholm BK, Mangia A, Paradiso A, Dirix LY, Van Laere SJ, Luider TM, Martens JW, Foekens JA and Umar A, J Natl Cancer Inst. 2014 Jan 7 (PubMed).

This study represents probably the best clinical proteomics data set obtained from laser microdisection samples. The starting material used in each analysis was approximately 4,000 human breast cancer epithelial cells removed from frozen tissue samples. The resulting set of spectra and identifications were surprisingly consistent,

Proteomics Informatics (BMSC-GA 4437)

Course Director

David Fenyö

Contact information

David@FenyoLab.org

http://fenyolab.org/presentations/Proteomics_Informatics_2014/