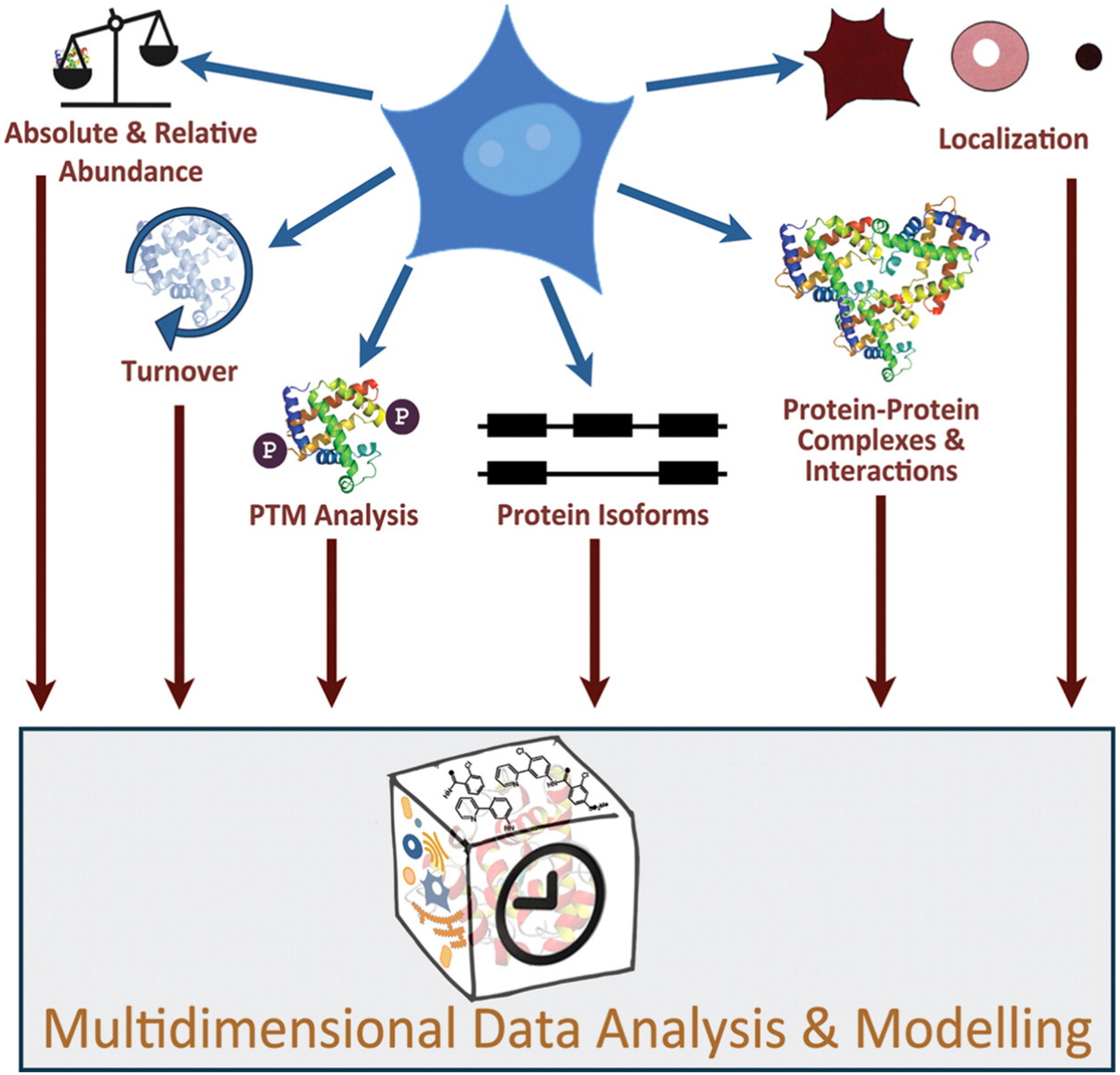
Introduction to Proteomics-Mass Spectrometry

Maria Person
Director, Proteomics Facility
MBB 1.420
pmaf@austin.utexas.edu
471-2895

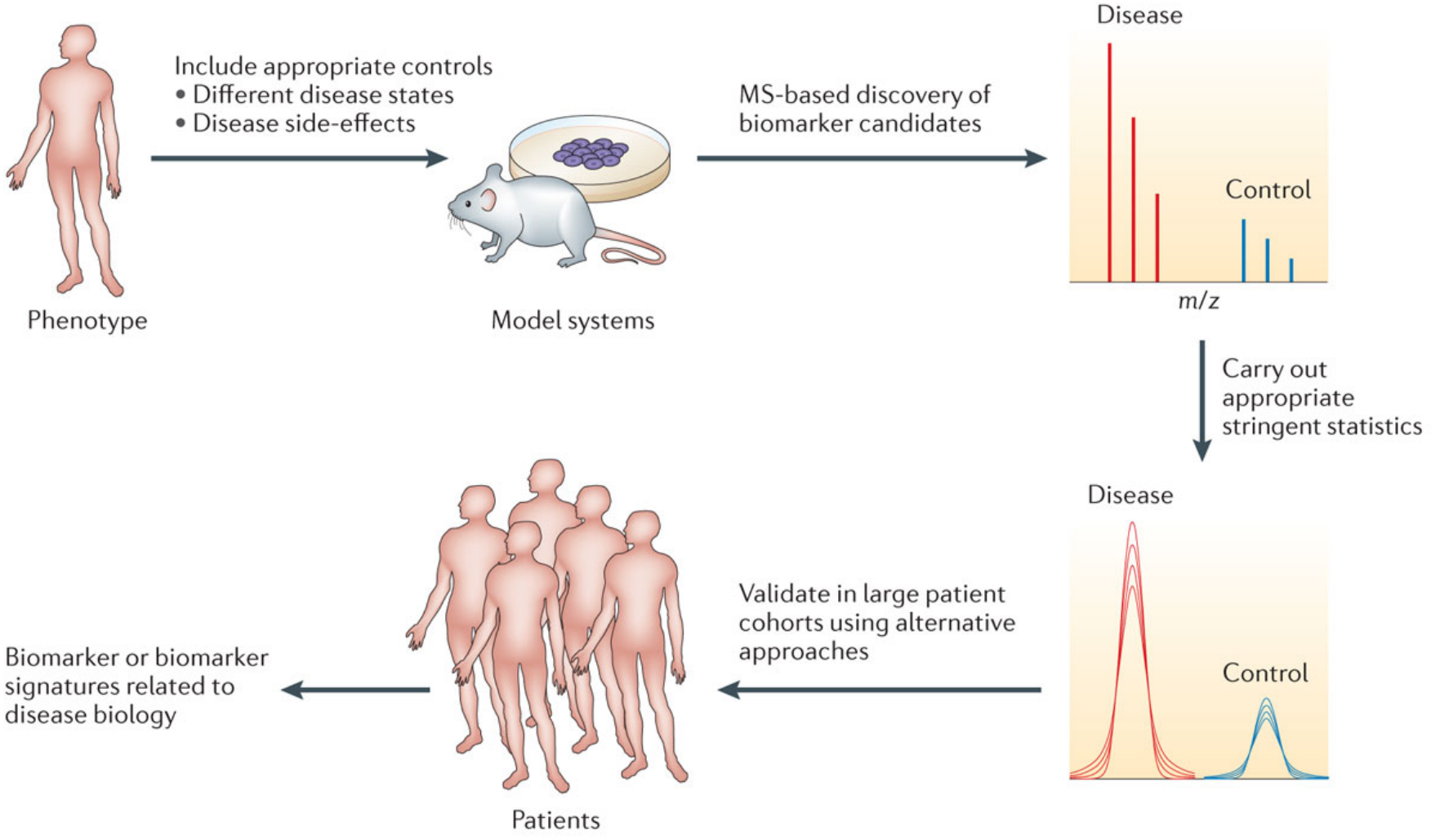
CCBB Short Course
April 28, 2016

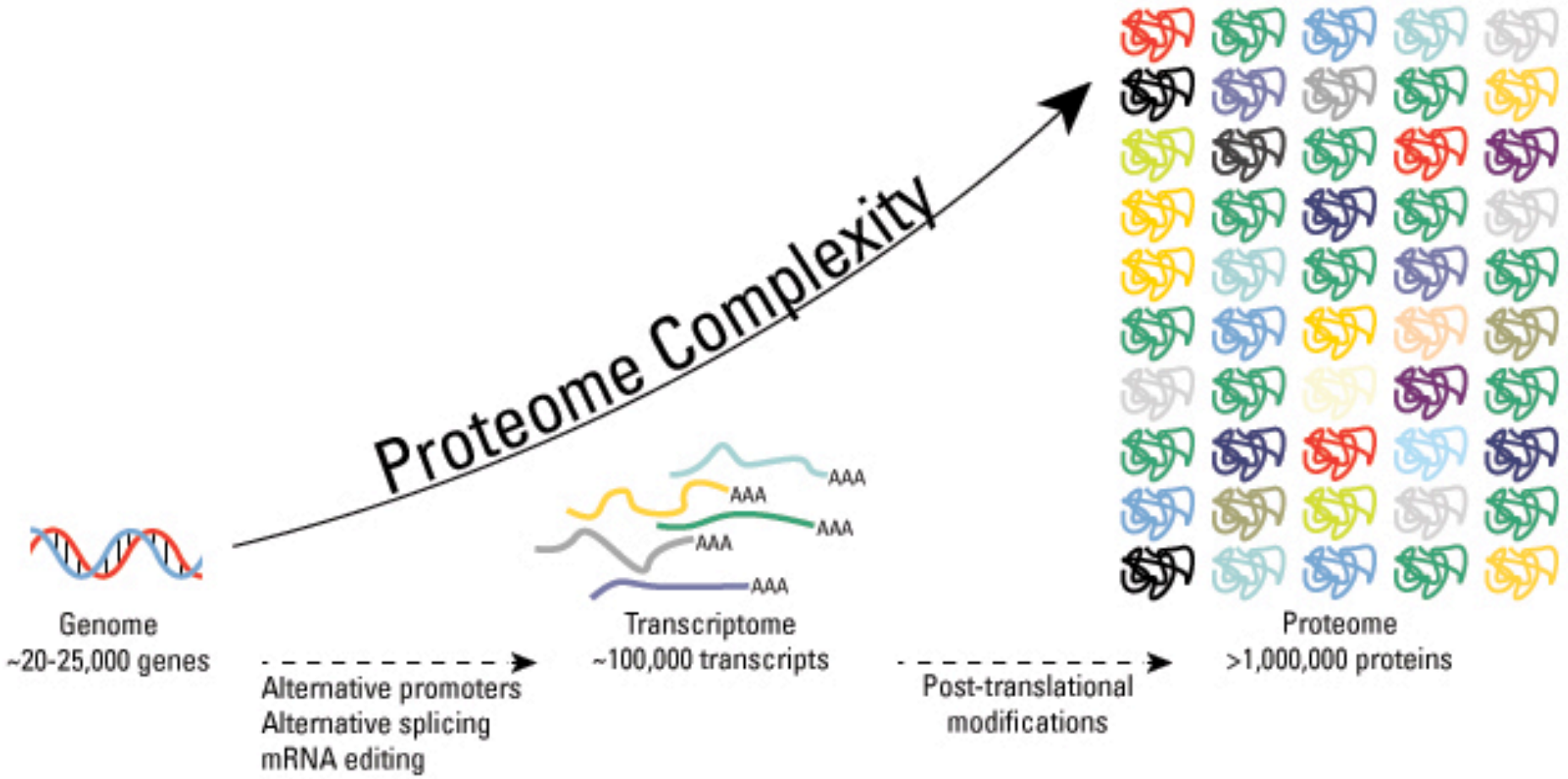
Outline

- Overview
- Protein/Peptide separation
- Mass spectrometry based protein identification with Scaffold 4
- Quantitative proteomics with Scaffold 4
- Post-translational modifications with Scaffold PTM



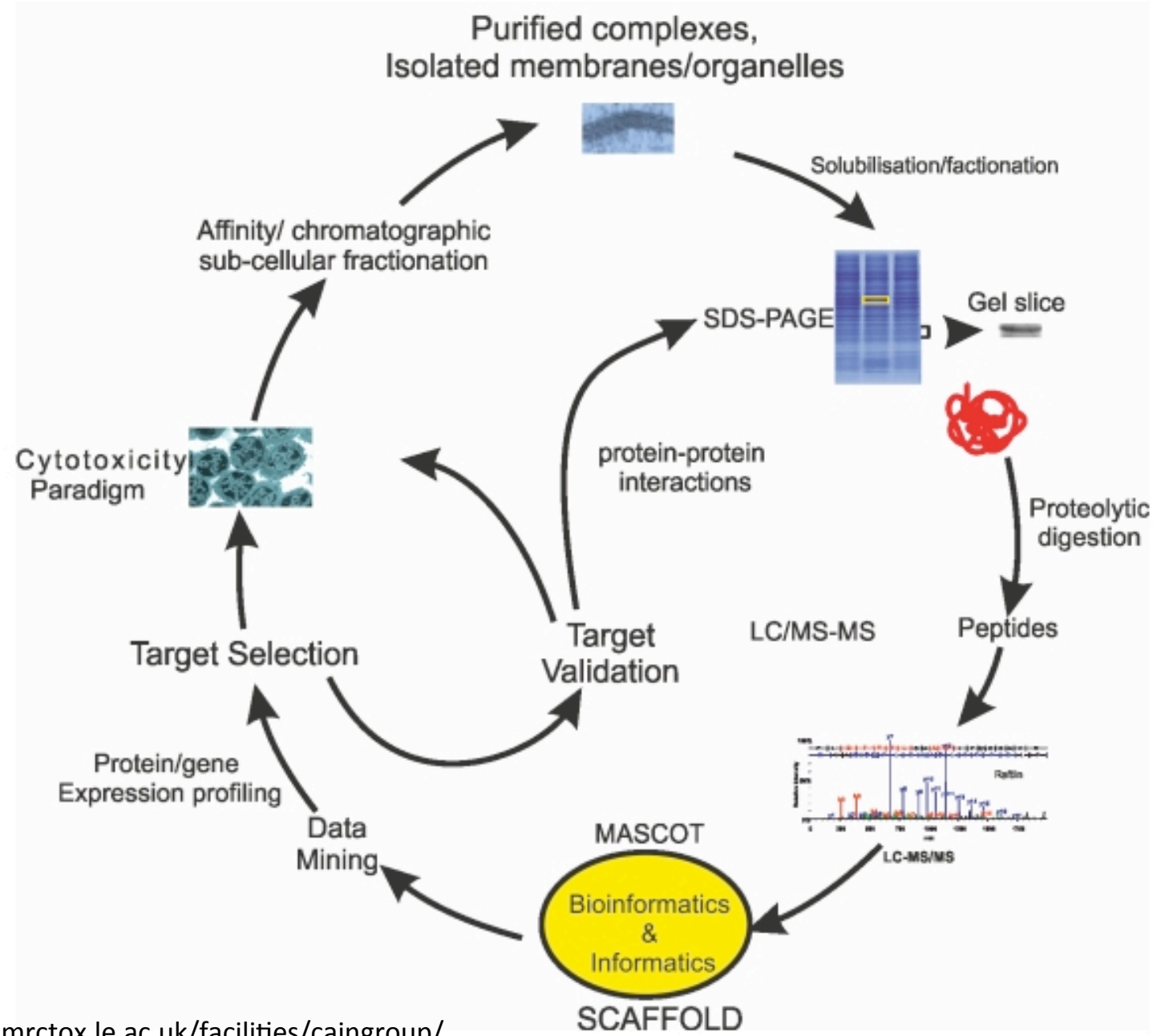
Lamond A I et al. Mol Cell Proteomics 2012;11:O112.017731


Phenotype
Include appropriate controls
• Different disease states
• Disease side-effects
Model systems
MS-based discovery of biomarker candidates
Disease
Control
m/z
Carry out appropriate stringent statistics
Disease
Control
Validate in large patient cohorts using alternative approaches
Patients
Biomarker or biomarker signatures related to disease biology



http://www.piercenet.com/browse.cfm?fldID=7CE3FCF5-0DA0-4378-A513-2E35E5E3B49B

Workflow to identify new cytotoxicity targets with quantitative proteomics



http://www.mrctox.le.ac.uk/facilities/caingroup/

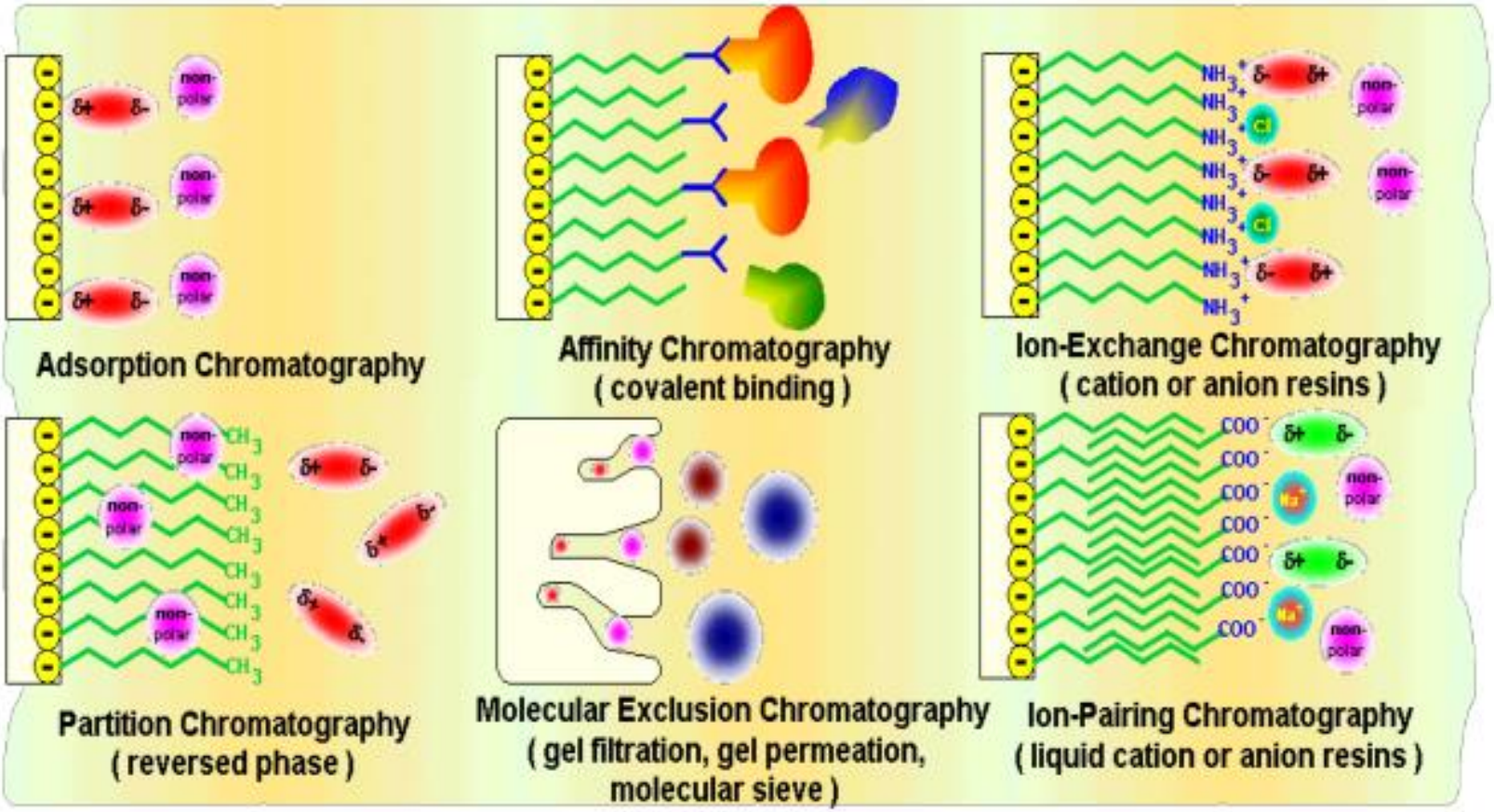
Protein and Peptide Sample Separation

Separation of complex samples

- Immobilized: gel electrophoresis, isoelectric focusing
- Liquid chromatography: Strong cation exchange (SCX), Reversed phase (RP), HILIC, WCX, Affinity chromatography

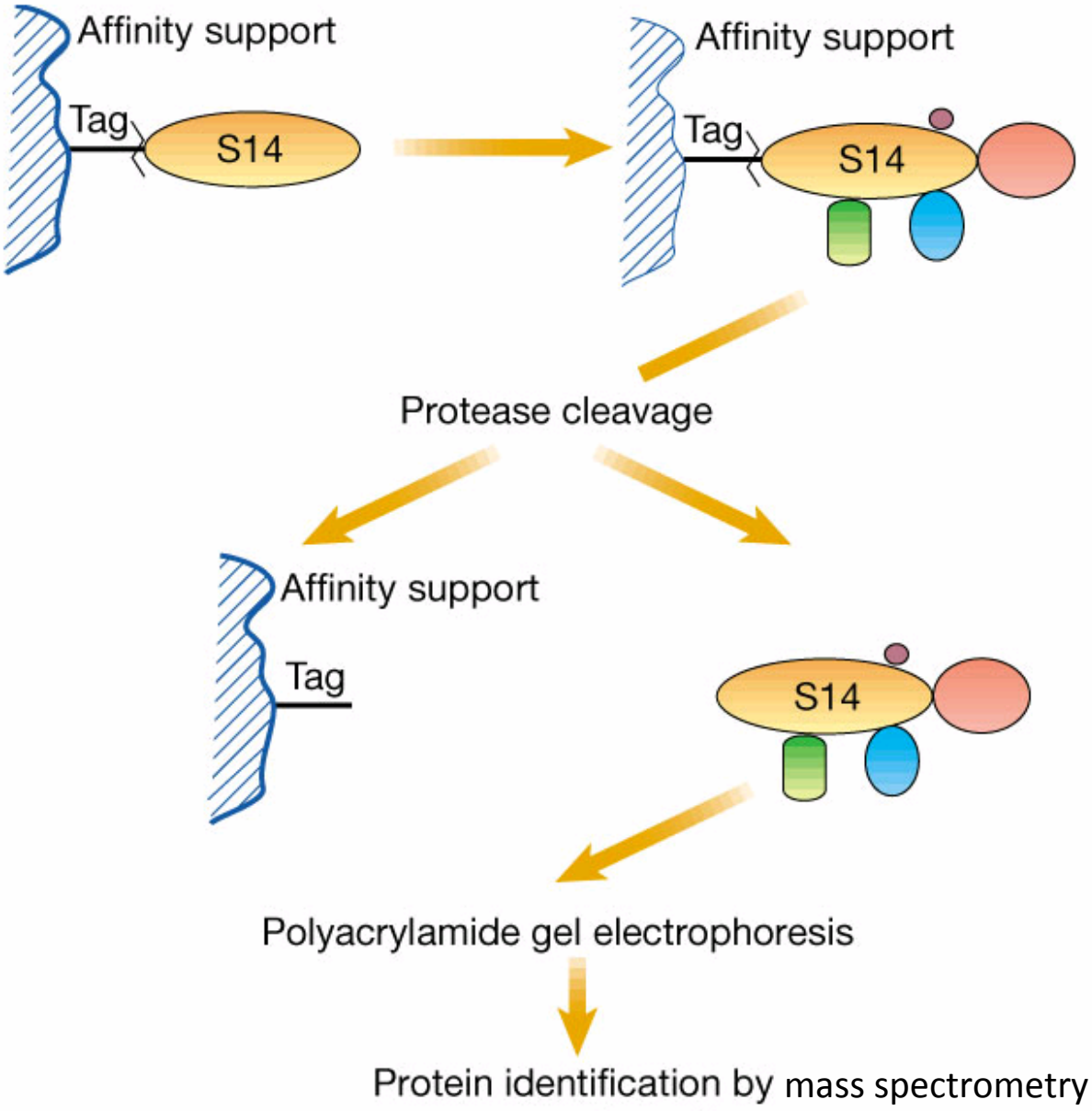
Methods combined for 2D separation: MudPIT (SCX-RPLC of peptides), 2DGE, GeLC (1D gel protein RPLC peptides)

Types of Chromatography



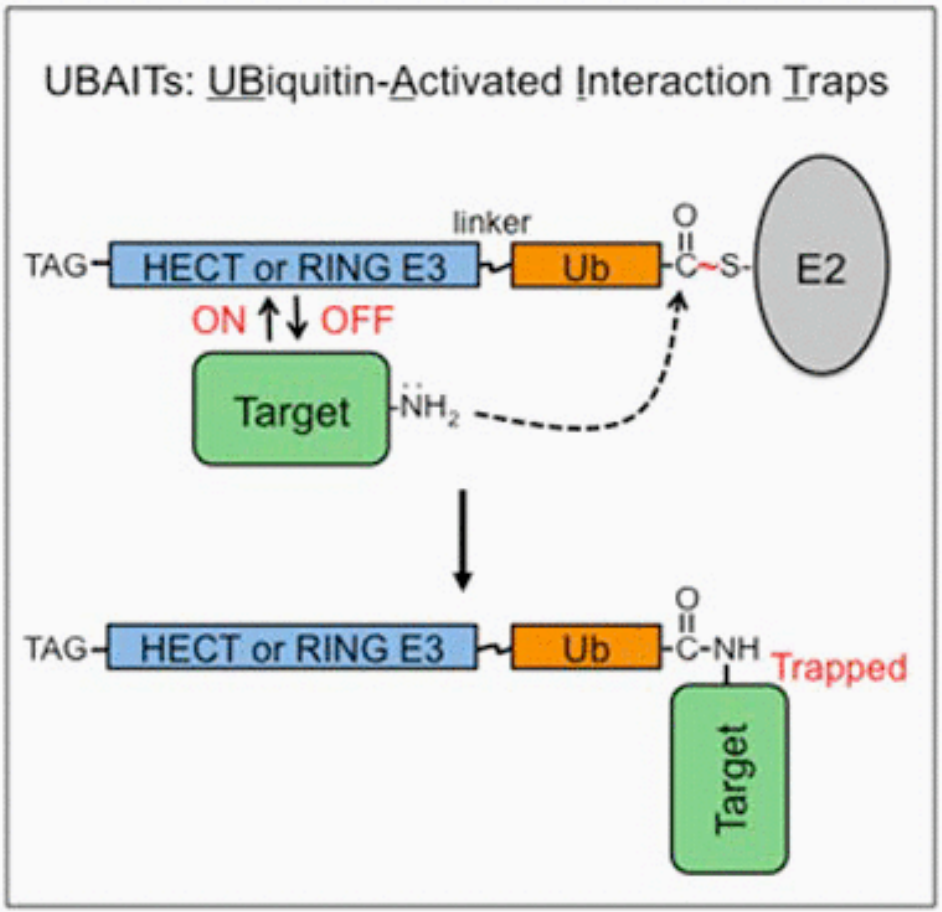
From Nina Salamah, Column Chromatography, http://image2.slideserve.com/5076567/types-of-chromatography-n.jpg

Identify binding partners to determine protein function



Pandey and Mann, *Nature* 405 *6788* 837 - 846 (2000)

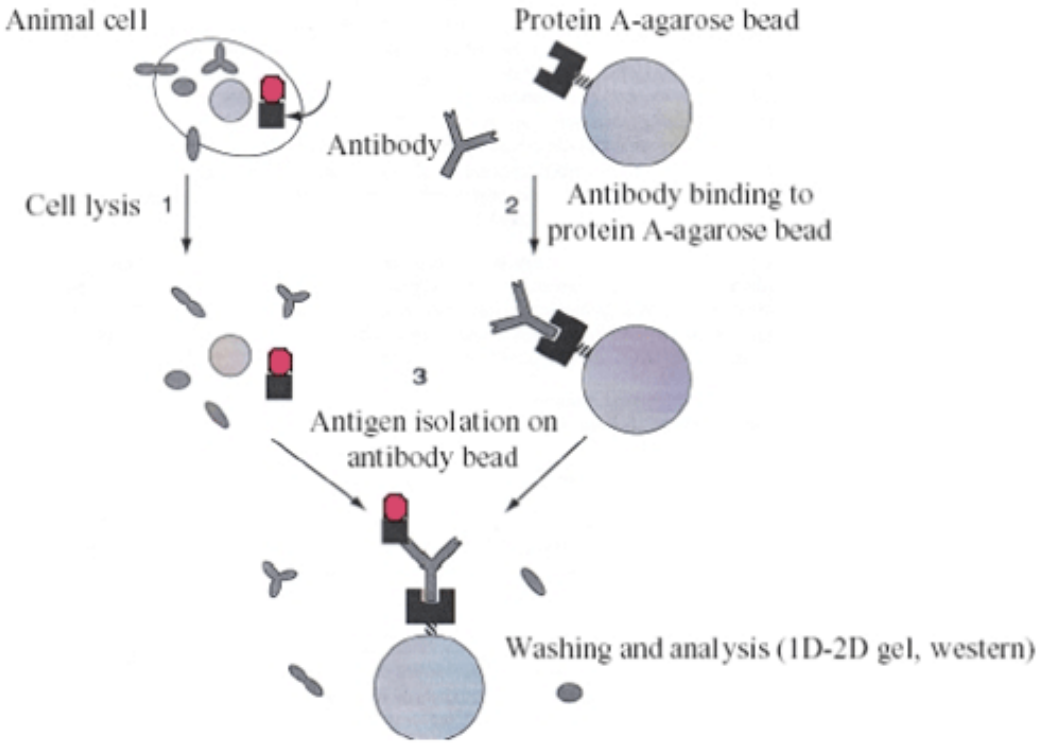
Huibregtse lab develops UBAIT method



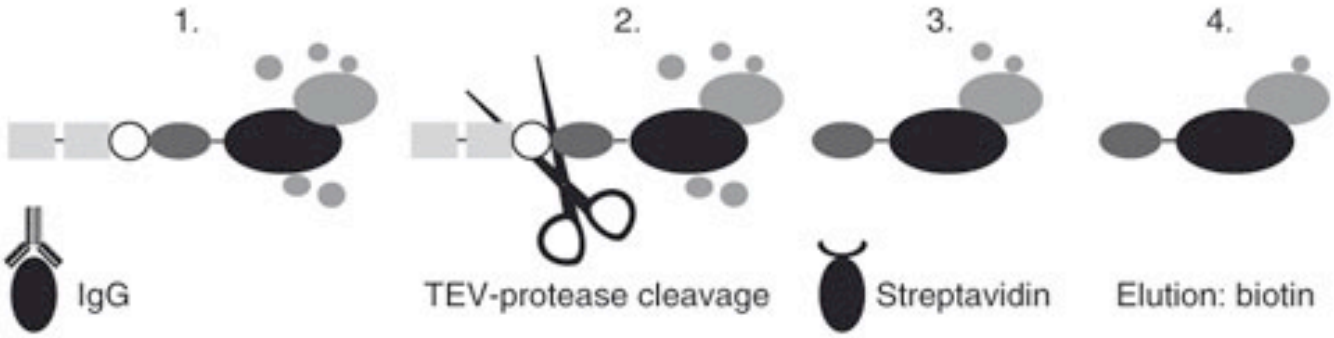
LC-MS/MS identifies known and novel E3 ligase interacting proteins from transient interactions

EMBO Rep. 2015 Dec;16(12):1699-712. doi: 10.15252/embr.201540620. Epub 2015 Oct 27.

Co-Immunoprecipitation: Protein specific antibodies

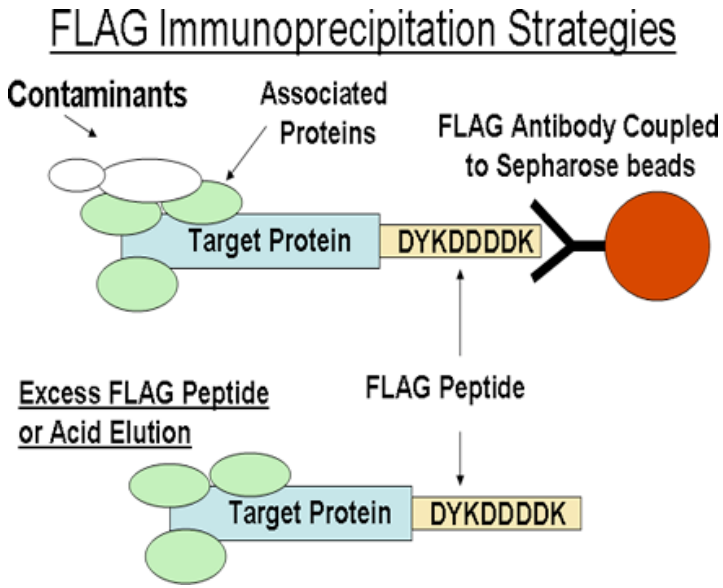


Tandem Affinity Purification TAP tag



Courtesy of Mark Bedford

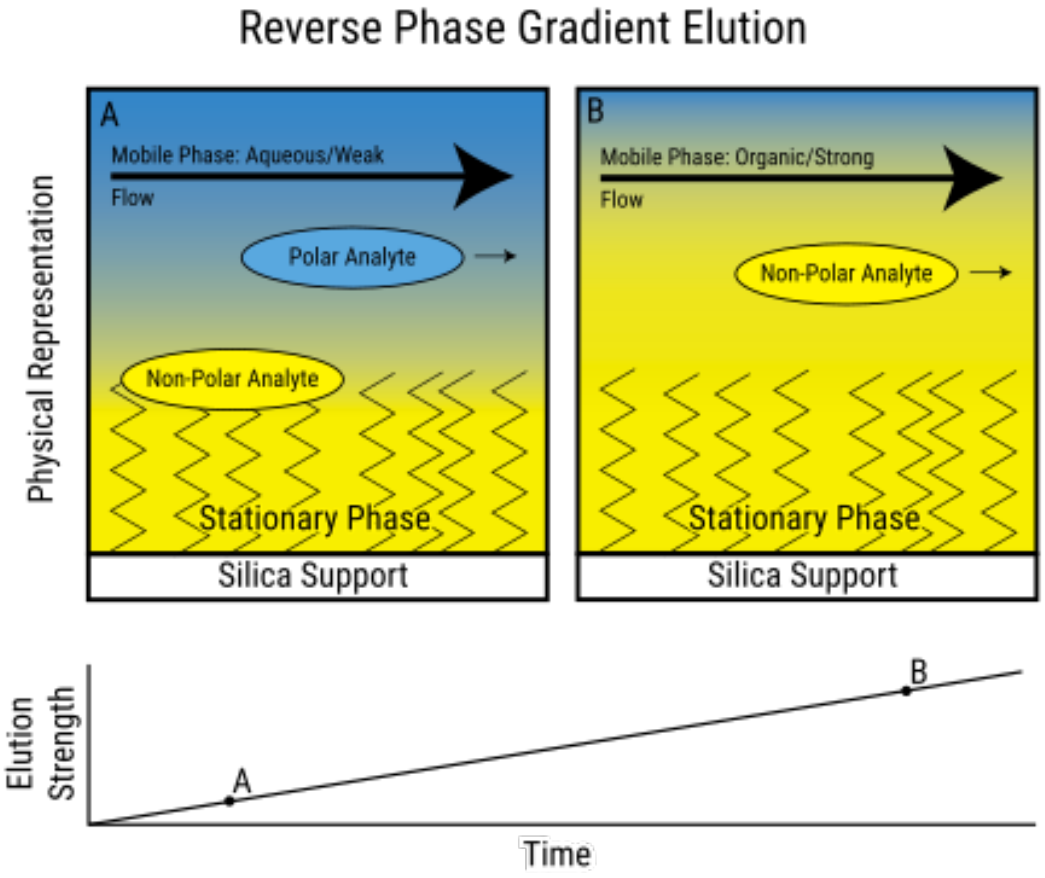
Most common epitope tags are:
His-tag
Flag-tag
V5-tag
Myc-tag
HA-tag



Problems – antibody cross-reactivity.

Courtesy of Mark Bedford

[Video of RPLC]



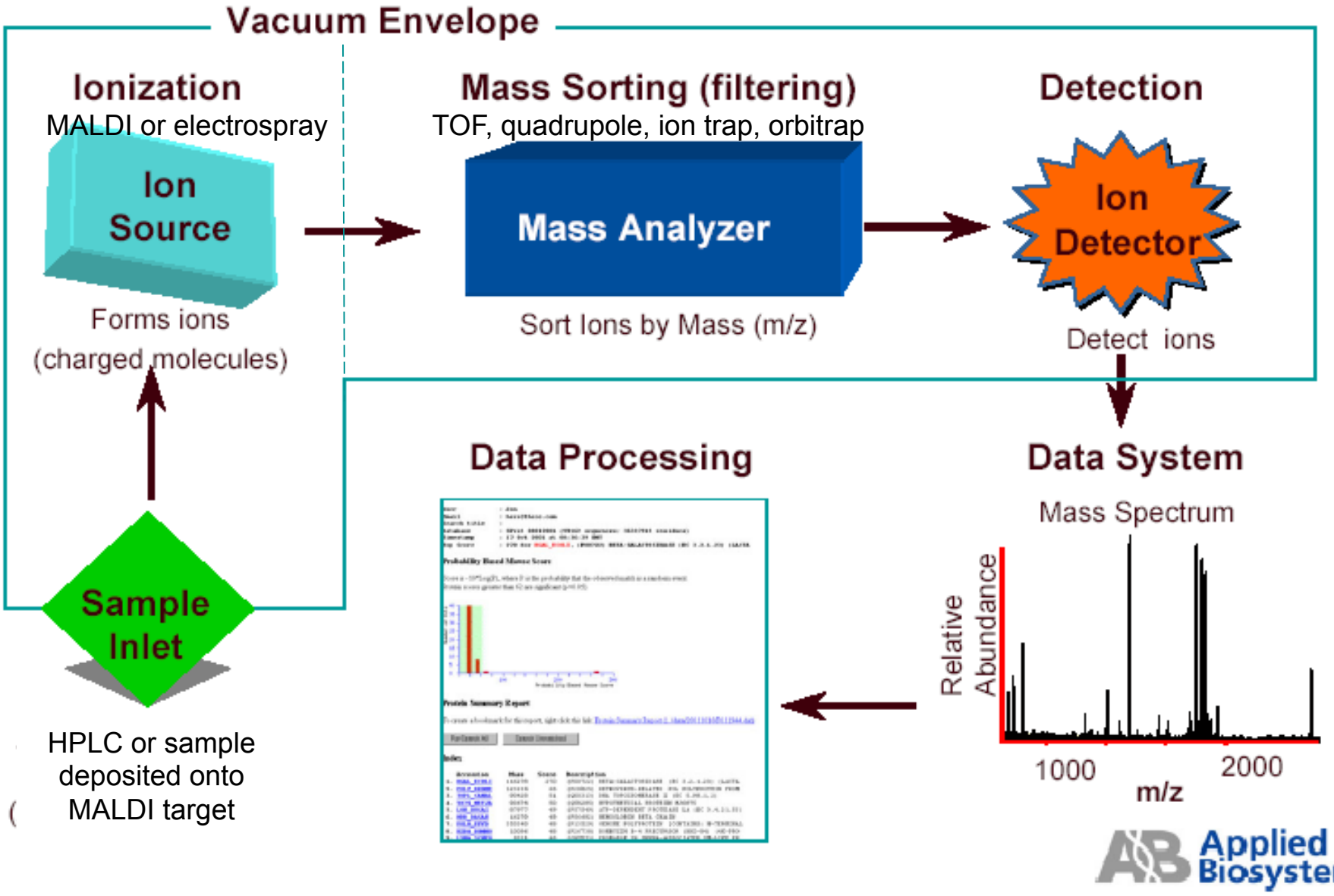
Proteomics Sample Preparation

Mass spectrometry requires buffer free samples:

- Run a gel, then can use in-gel digest to remove unwanted buffer components
- TCA or acetone precipitation and wash lysate
- Ziptip / membrane centrifugation / dialysis / Sep Pack to remove salts, esp. Na or K or phosphate
- Avoid use of polymers and detergents, i.e. Triton-X, NP-40, SDS, glycerol; use urea and mass spec friendly detergents instead or remove with Pierce detergent removal kit
- separate and purify components—HPLC

Mass Spectrometry Based Protein Identification

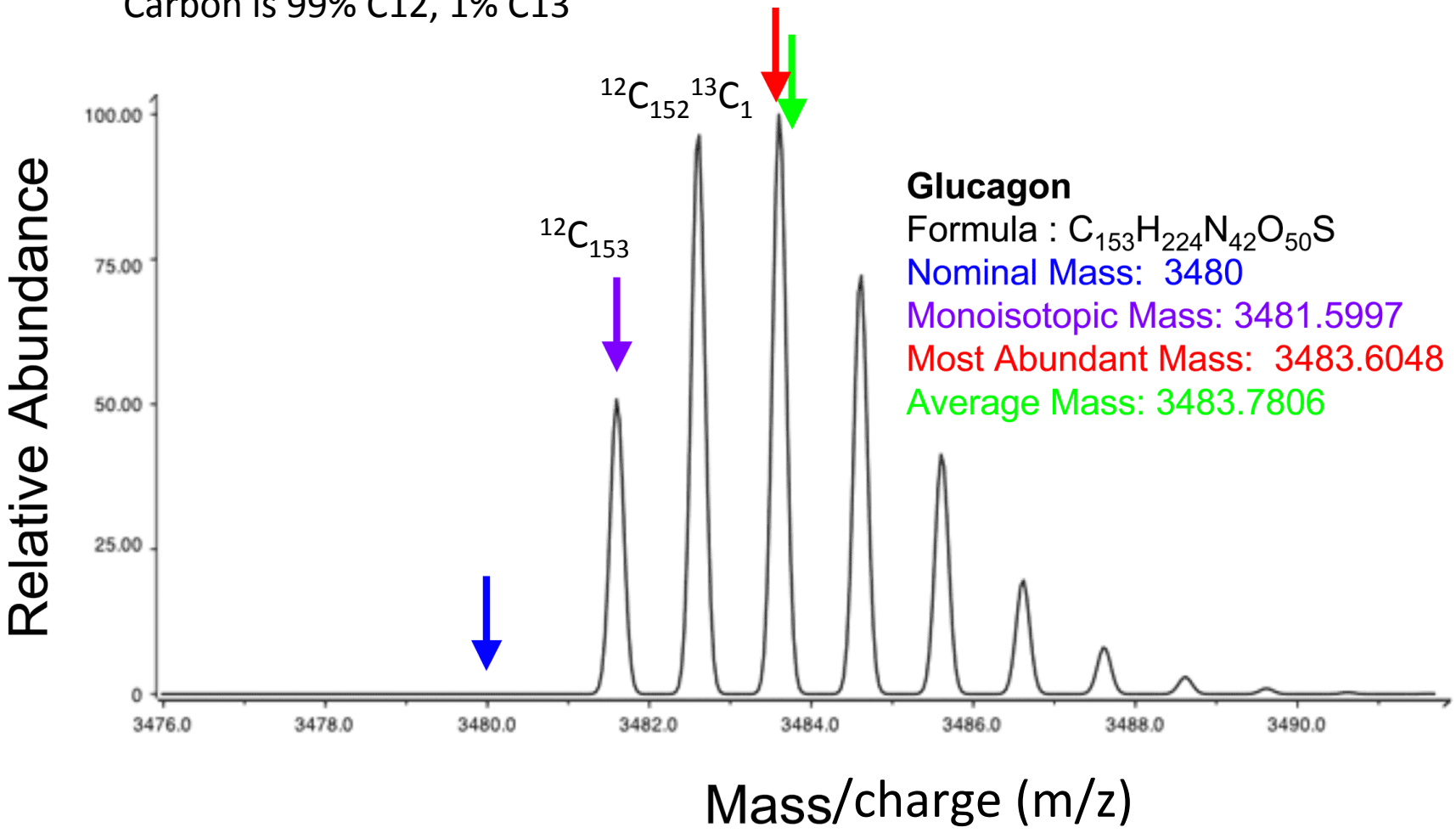
Basic Components of a Mass Spectrometer



Applied Biosystems

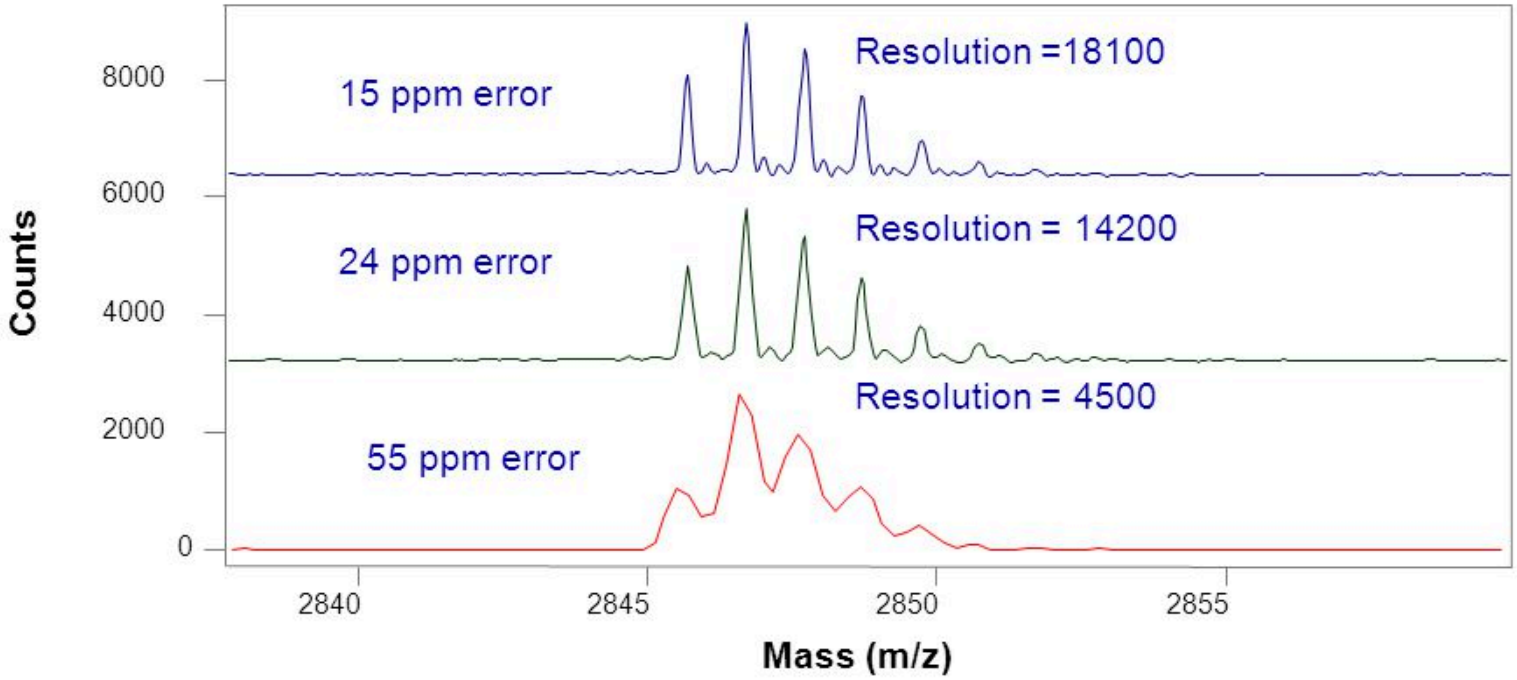
Peptide mass spectrum

Monoisotopic peak has C12,H1, N14, O16
Carbon is 99% C12, 1% C13



Mass measurement accuracy depends on resolution

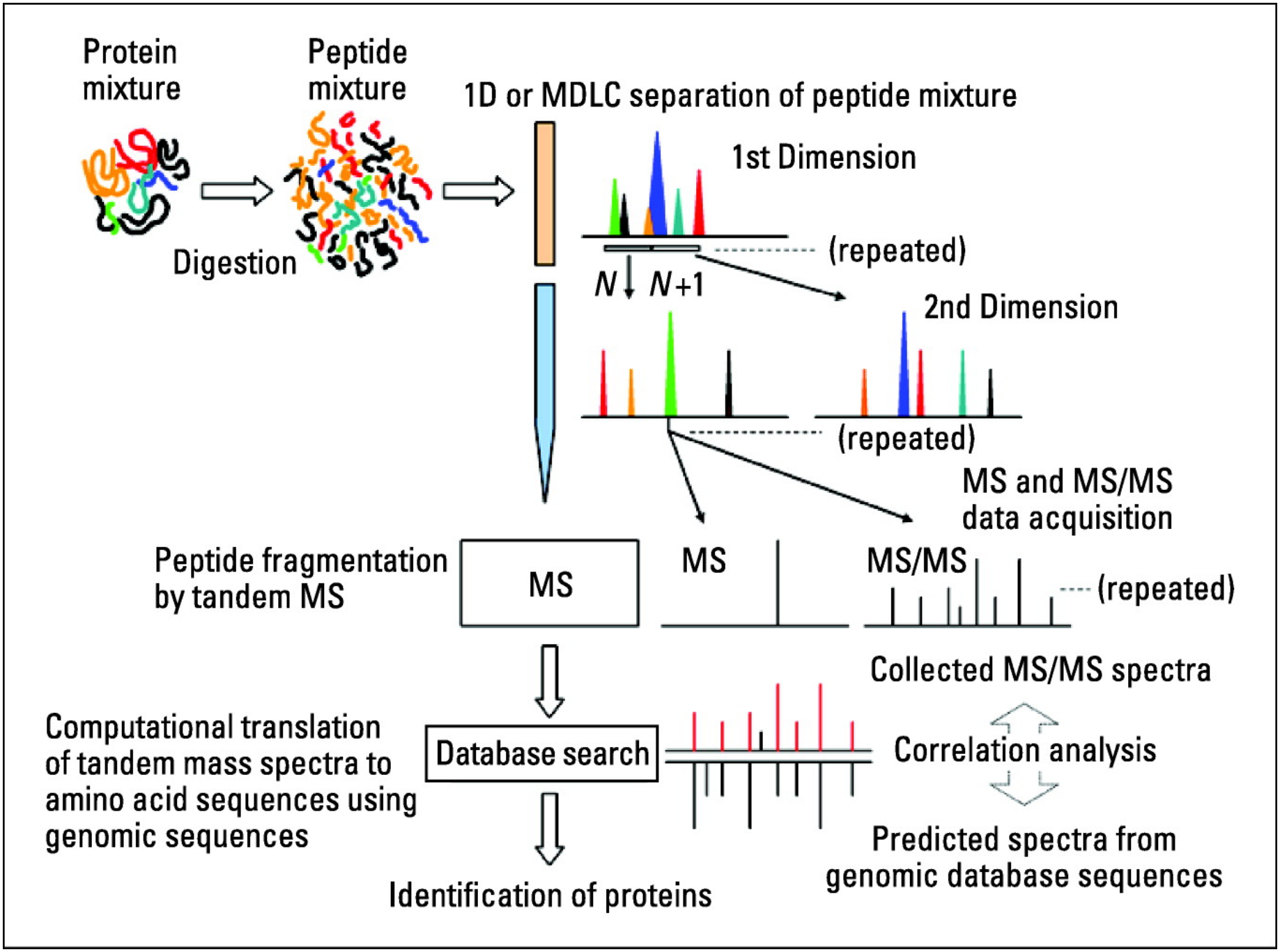
High resolution means better mass accuracy



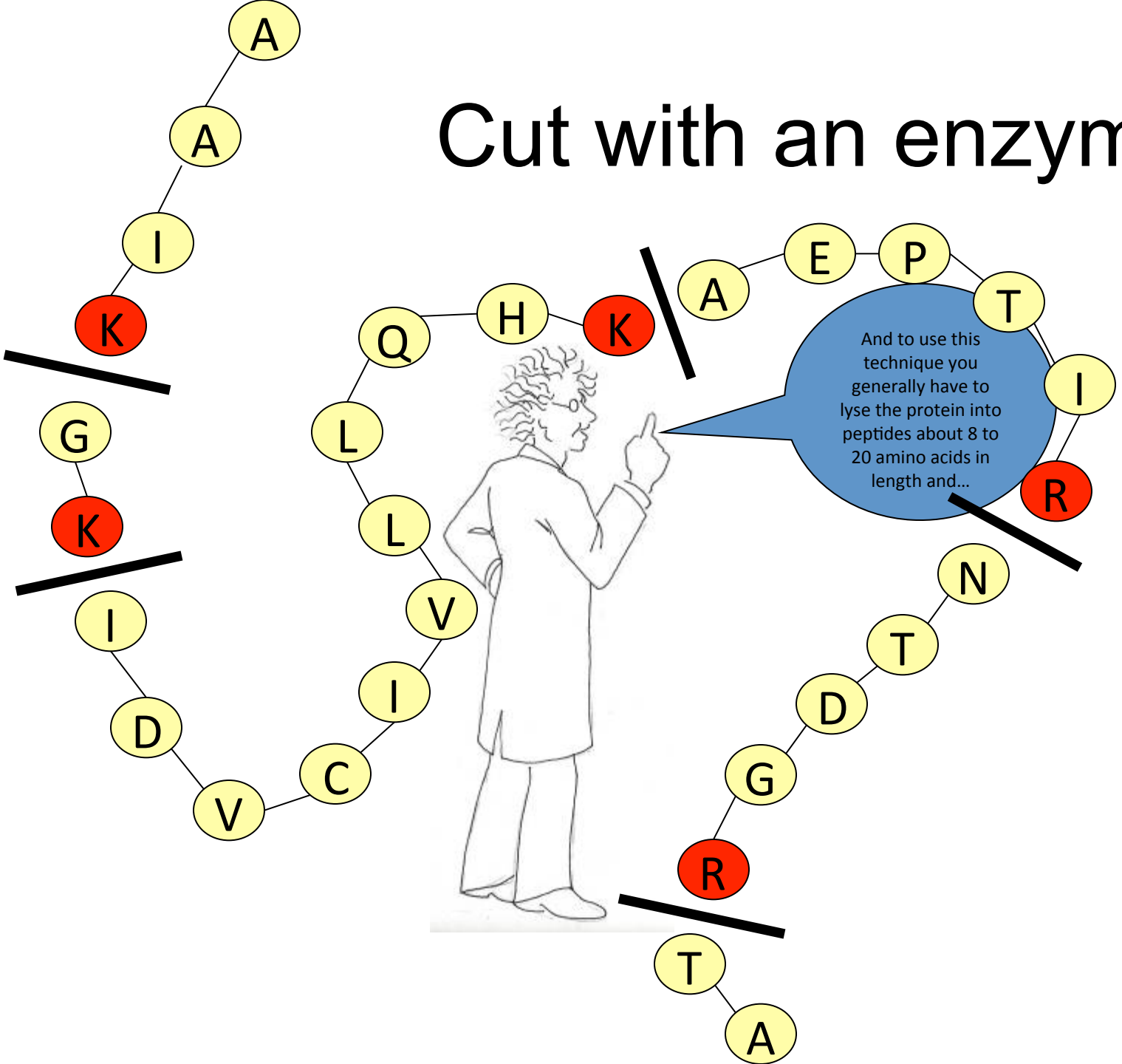
High mass accuracy improves selectivity in database searches
High resolution separates peptides by mass better in complex samples
Orbitrap Fusion operates at 10 ppm and 120,000 resolution routinely

nanoflowUPLC-Orbitrap Fusion

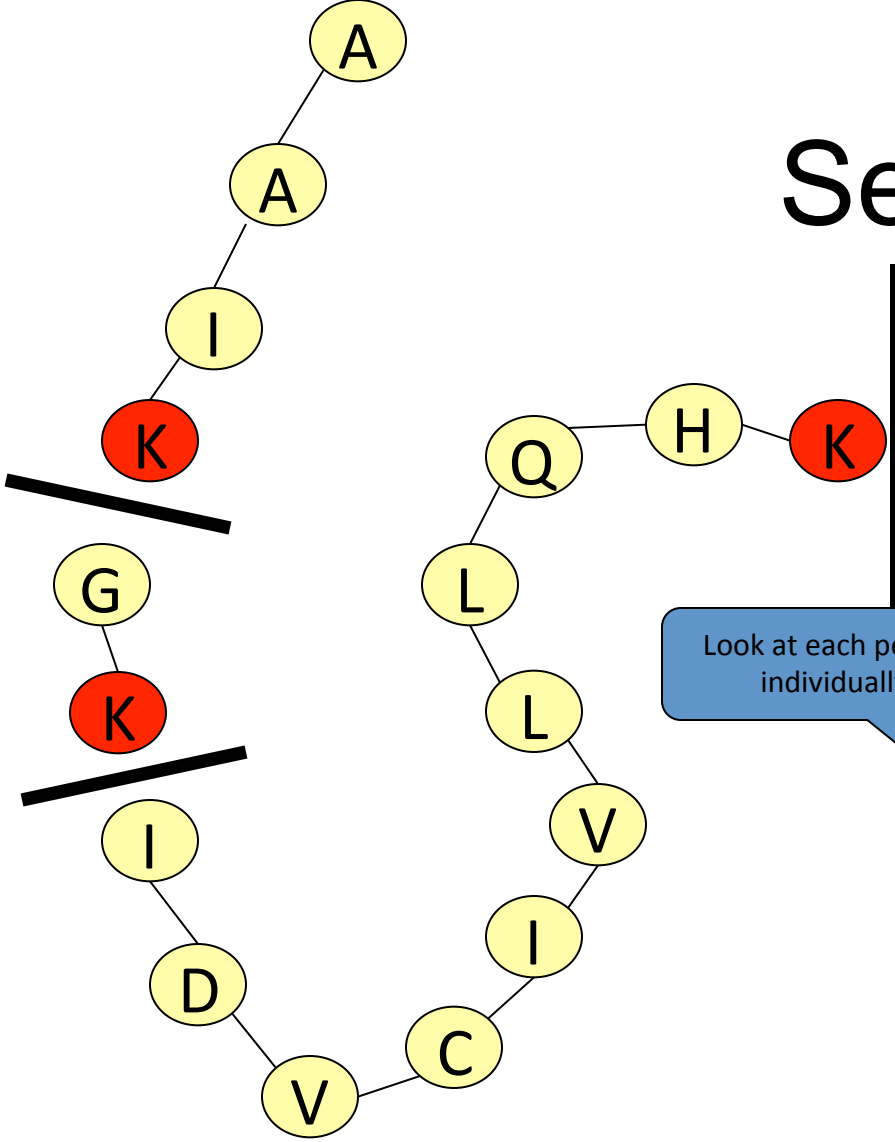
- UPLC—up to 800 bar for better separation of peptides using reversed phase liquid chromatography at low flow
- FT-MS—resolution max 450,000 and low ppm mass accuracy improves peptide ID confidence
- Sensitivity—1 fmol digest BSA standard protein ID, ~10 fmol for spiked in proteins in cell lysate
- Complexity—10^3 proteins and 10^4 peptides identified in LC-MS/MS runs on lysates
- Quantitation—label (TMT, SILAC) and label free methods (peak area and intensity, spectral counting); linear dynamic range 3-4 orders of magnitude
- CID, HCD and ETD fragmentation choices for PTMs
- Orbitrap Fusion ion path

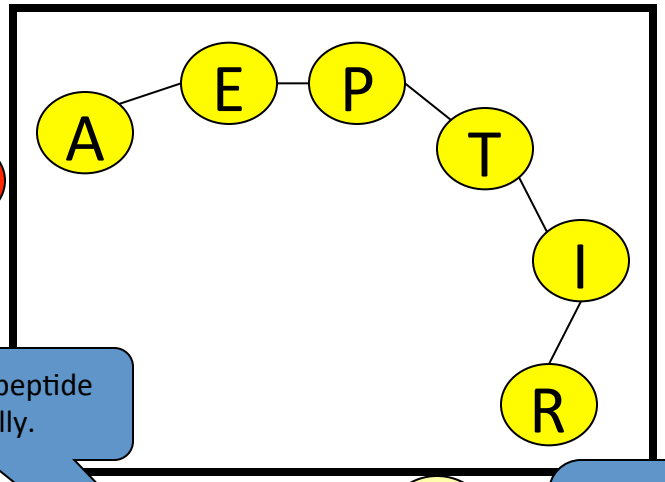

Protein mixture
Peptide mixture
Digestion
1D or MDLC separation of peptide mixture
1st Dimension
(repeated)
N
N+1
2nd Dimension
(repeated)
MS and MS/MS data acquisition
Peptide fragmentation by tandem MS
MS
MS
MS/MS
(repeated)
Collected MS/MS spectra
Computational translation of tandem mass spectra to amino acid sequences using genomic sequences
Database search
Correlation analysis
Predicted spectra from genomic database sequences
Identification of proteins

Cut with an enzyme



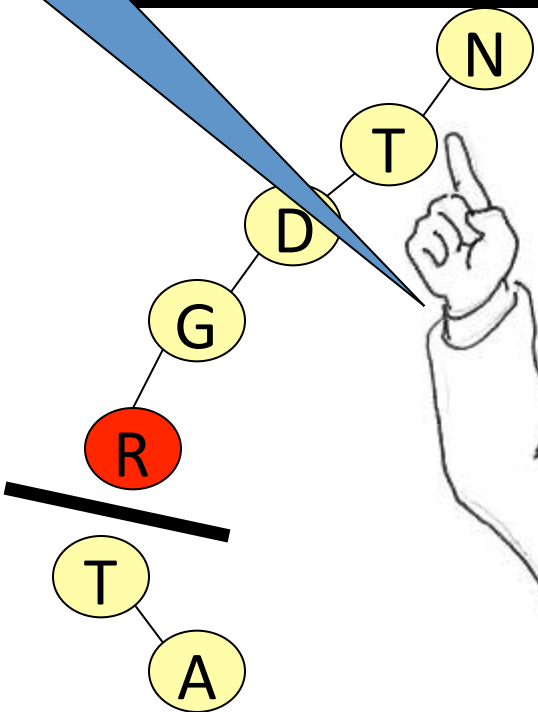
Select a peptide



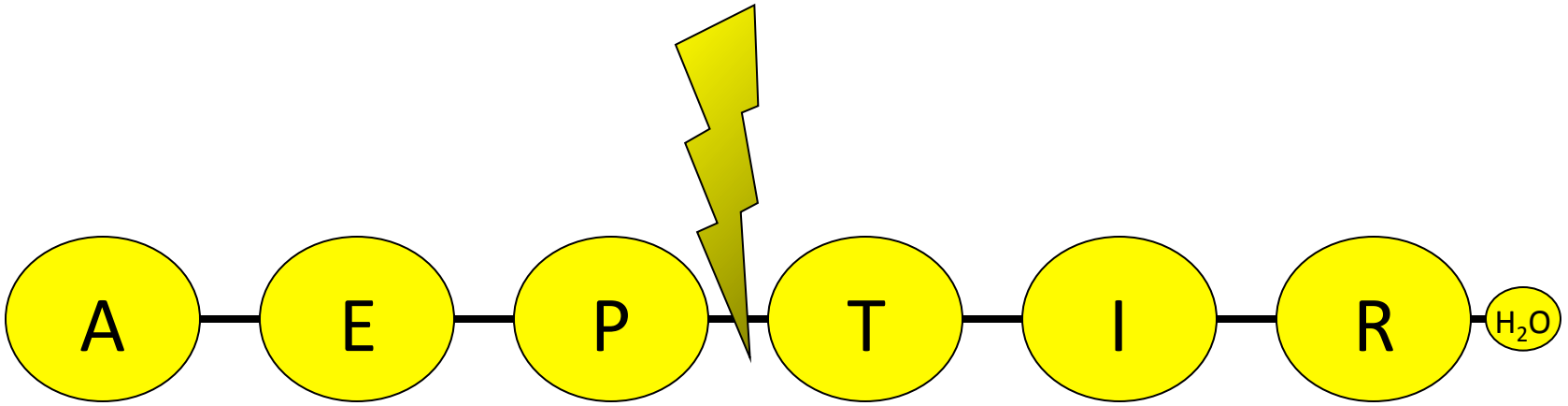


Look at each peptide individually.

We select the peptide by mass using the first half of the tandem mass spectrometer

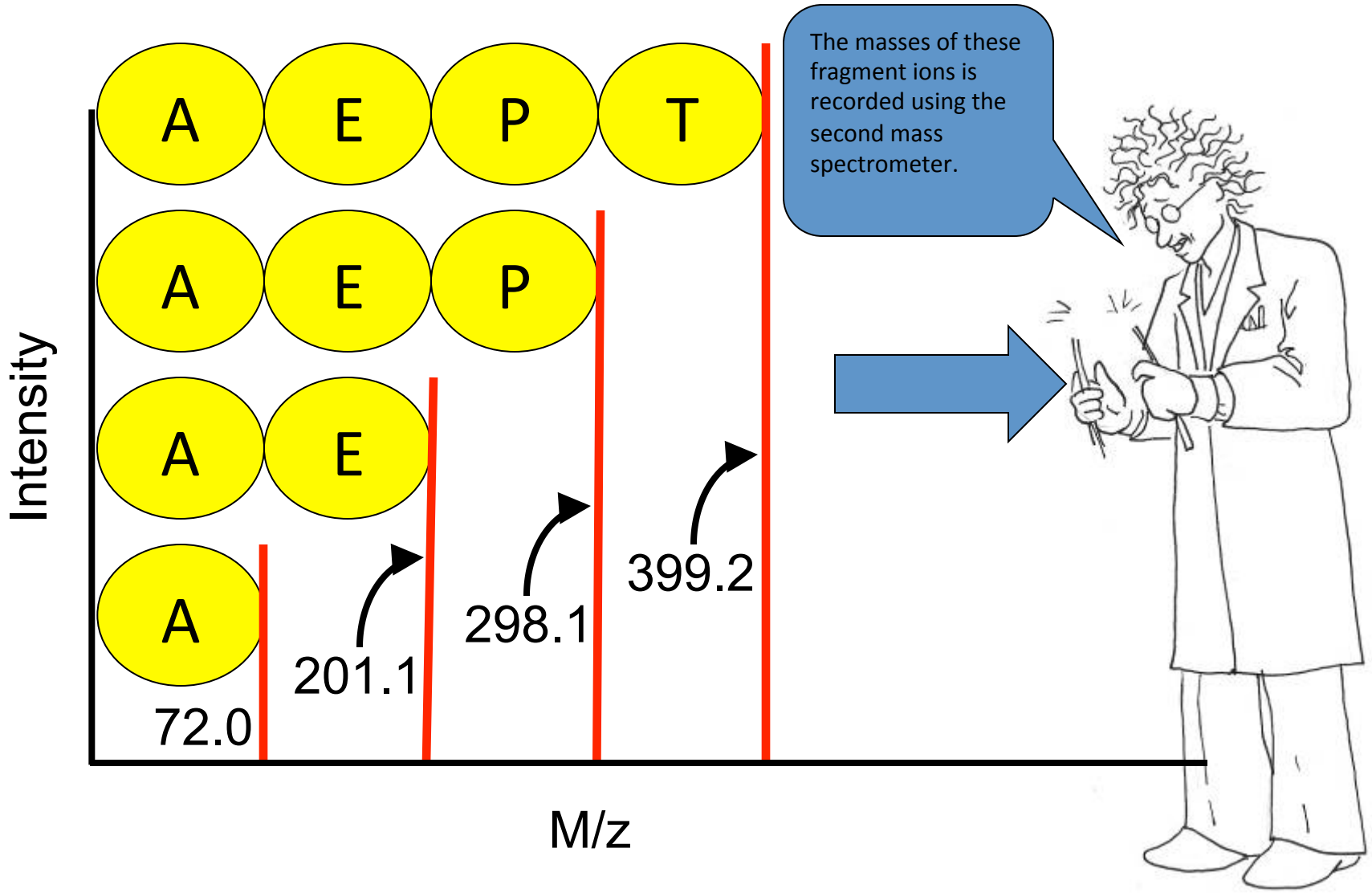


Impart energy in collision cell

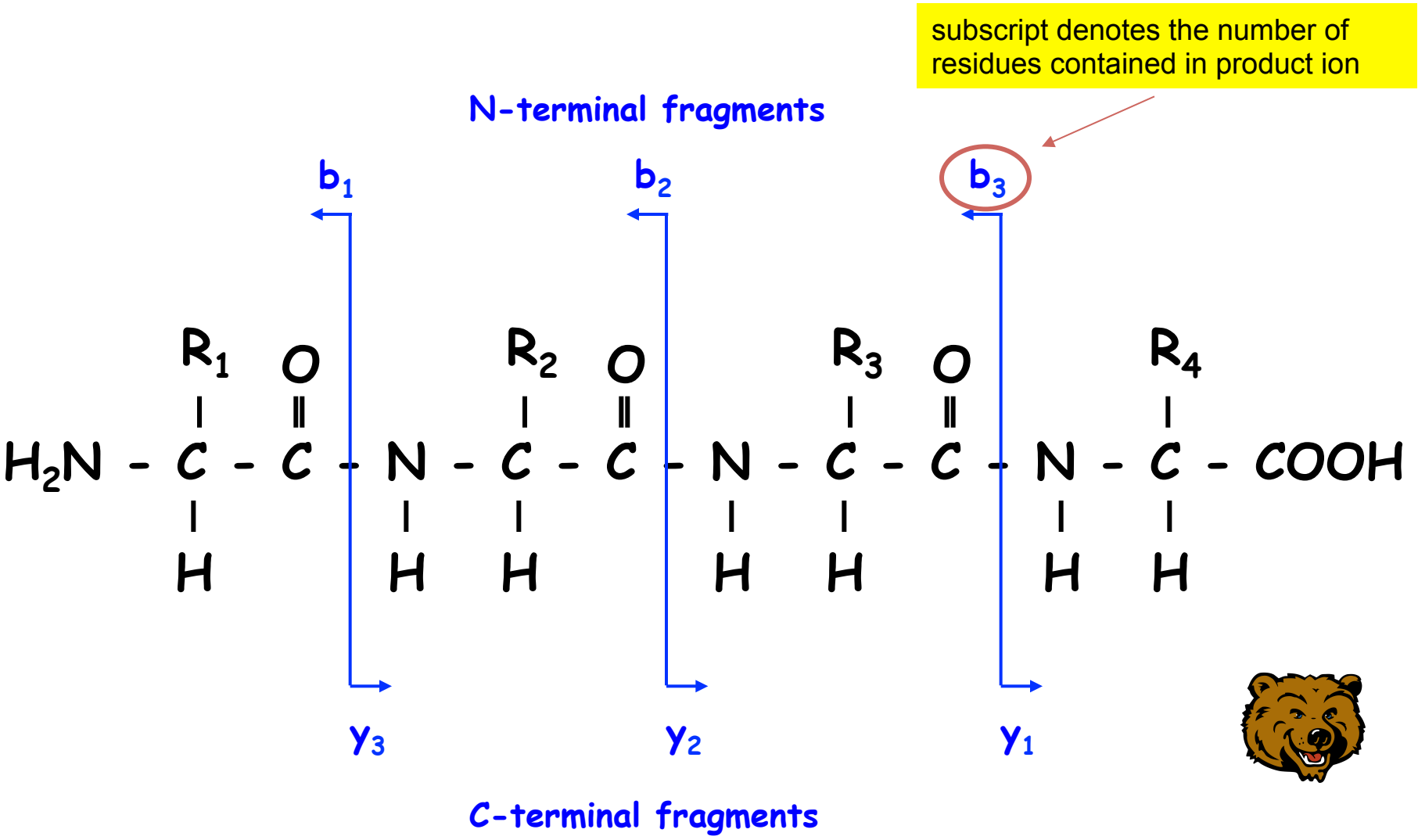


The mass spectrometer imparts energy into the peptide causing it to fragment at the peptide bonds between amino acids.

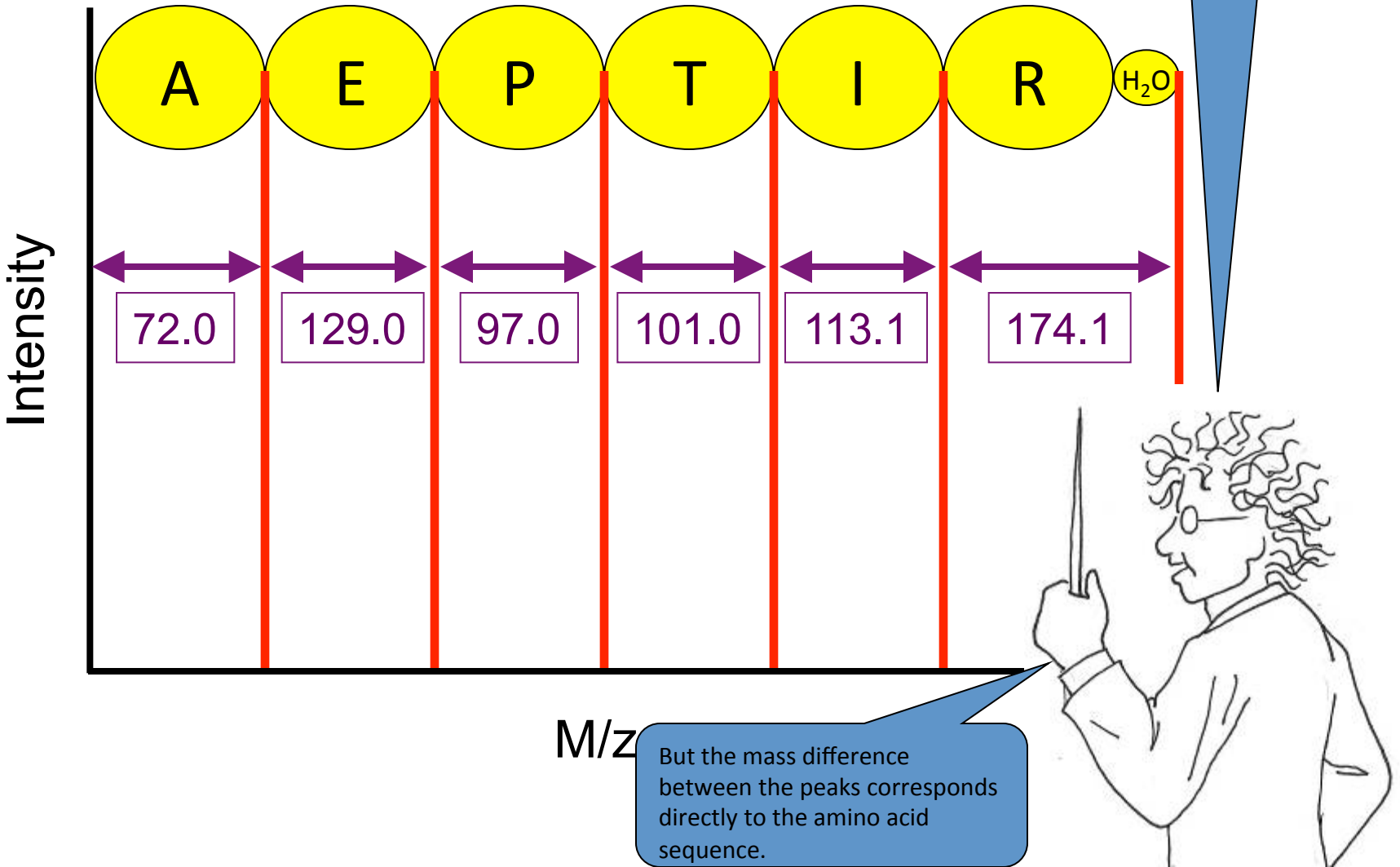
Measure mass of product ions



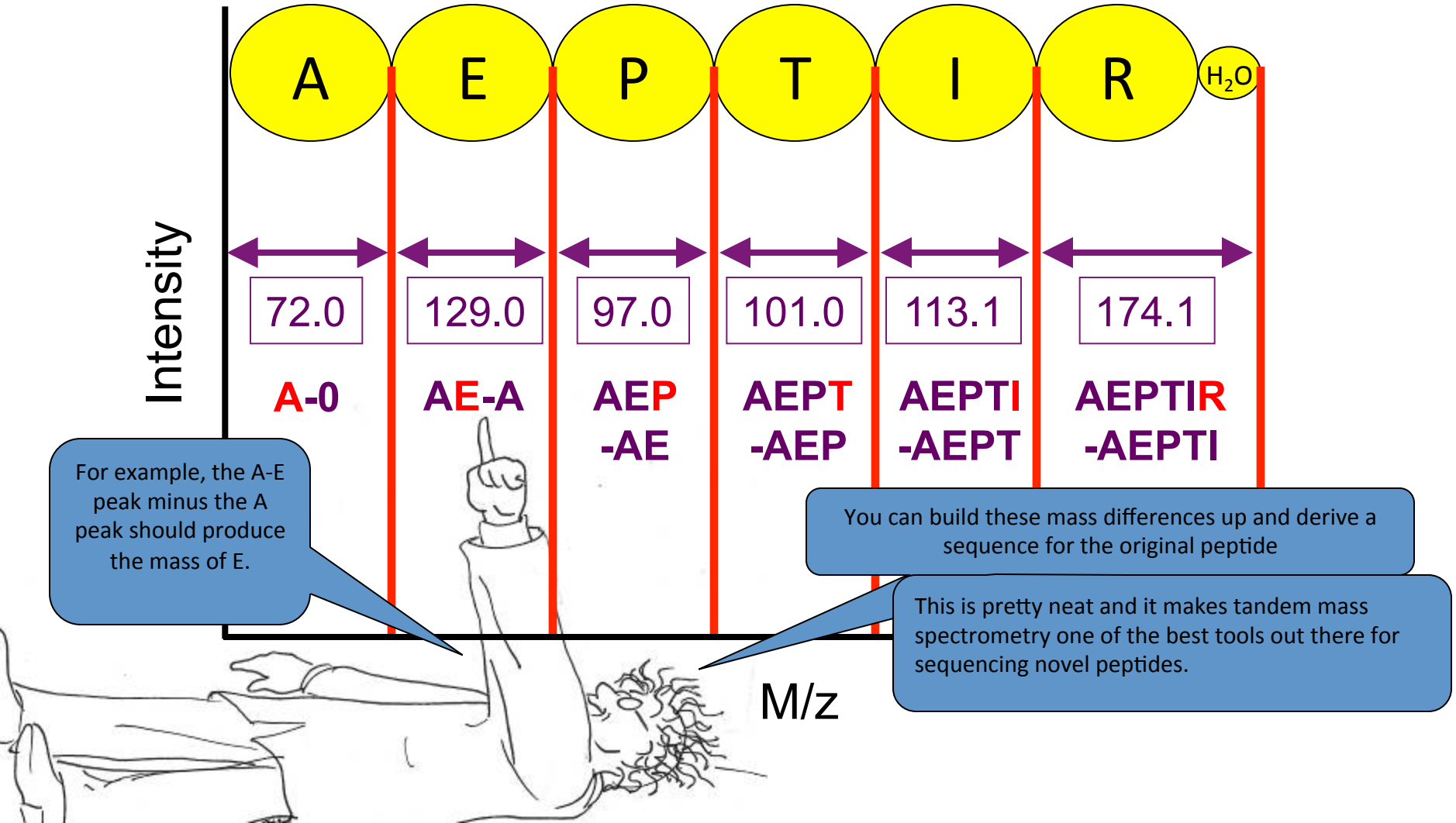
Nomenclature for MS Sequencing of Peptides



B-type Ions


These ions are commonly called B ions, based on nomenclature you don't really want to know about…
A
E
P
T
I
R
H_2O
Intensity
72.0
129.0
97.0
101.0
113.1
174.1
M/z
But the mass difference between the peaks corresponds directly to the amino acid sequence.

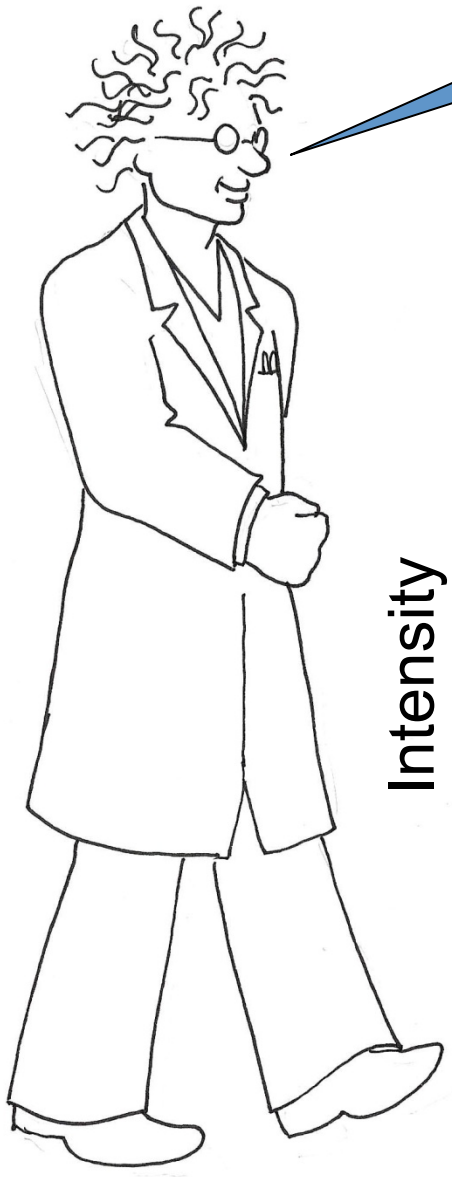
B-type Ions

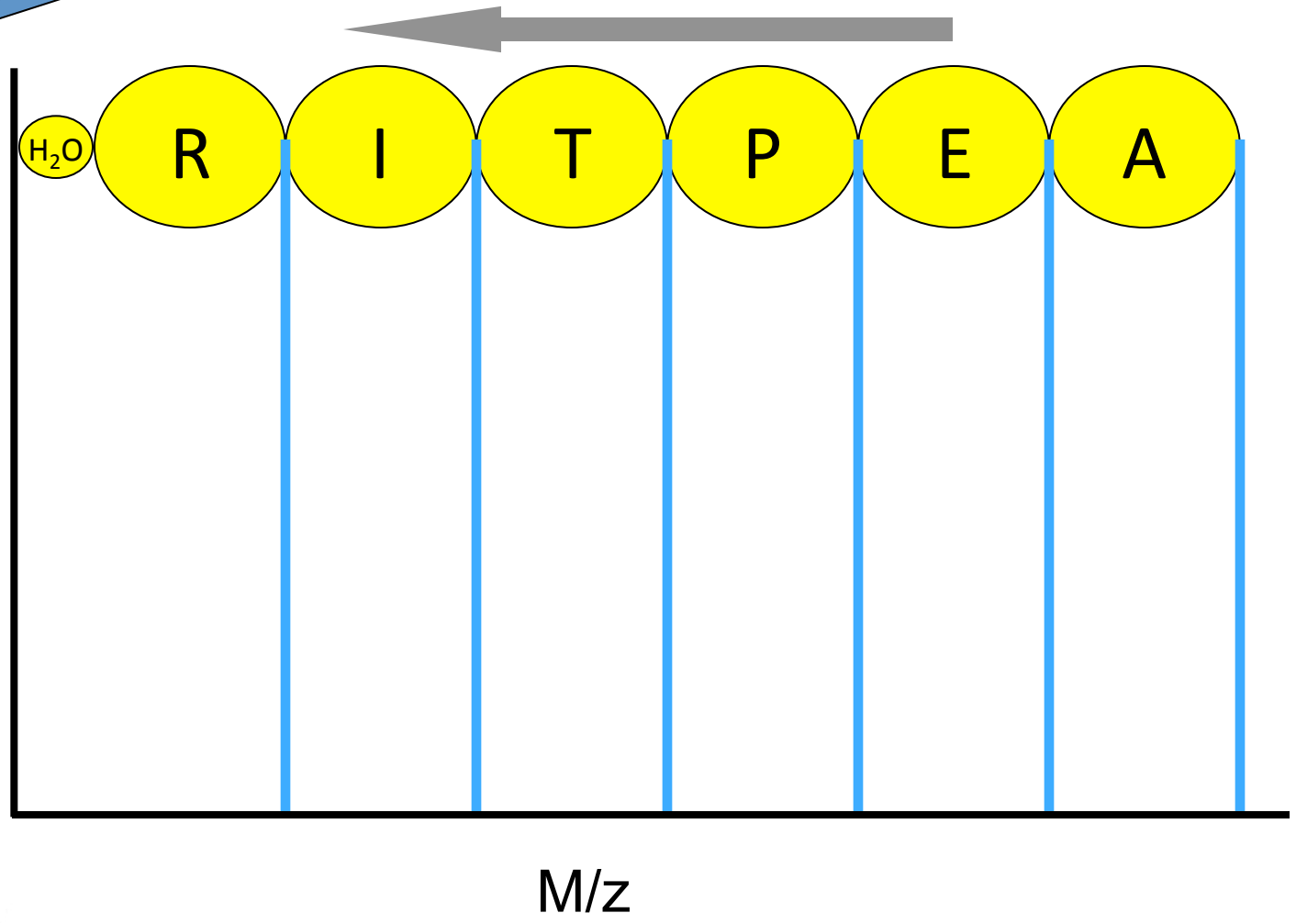

A
E
P
T
I
R
H_2O
Intensity
72.0
129.0
97.0
101.0
113.1
174.1
A-0
AE-A
AEP
-AE
AEPT
-AEP
AEPTI
-AEPT
AEPTIR
-AEPTI
M/z
For example, the A-E peak minus the A peak should produce the mass of E.
You can build these mass differences up and derive a sequence for the original peptide
This is pretty neat and it makes tandem mass spectrometry one of the best tools out there for sequencing novel peptides.

Y-type Ions

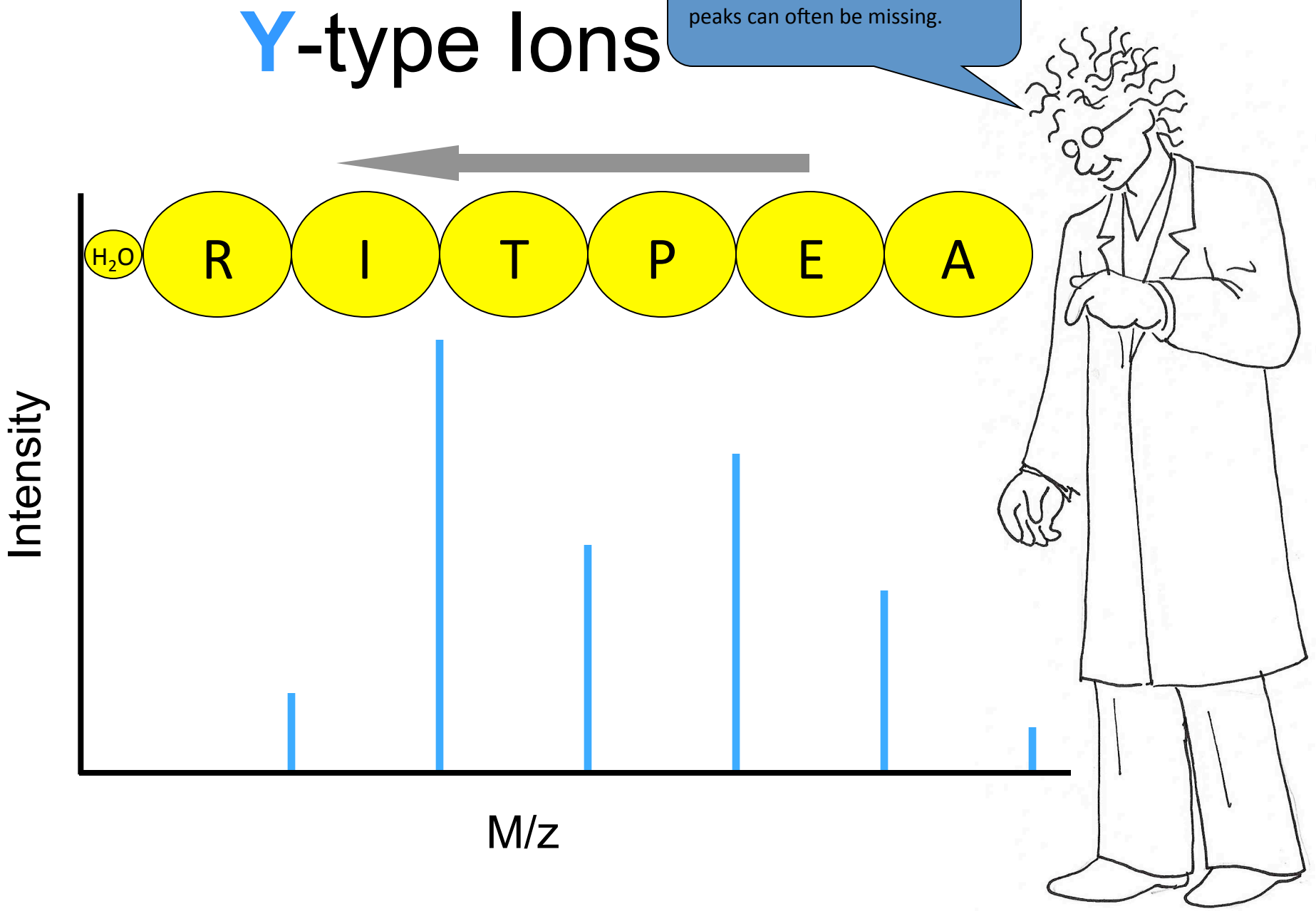
… The second half are represented as Y ions that sequence backwards.

And, unfortunately, this is the real world, so…



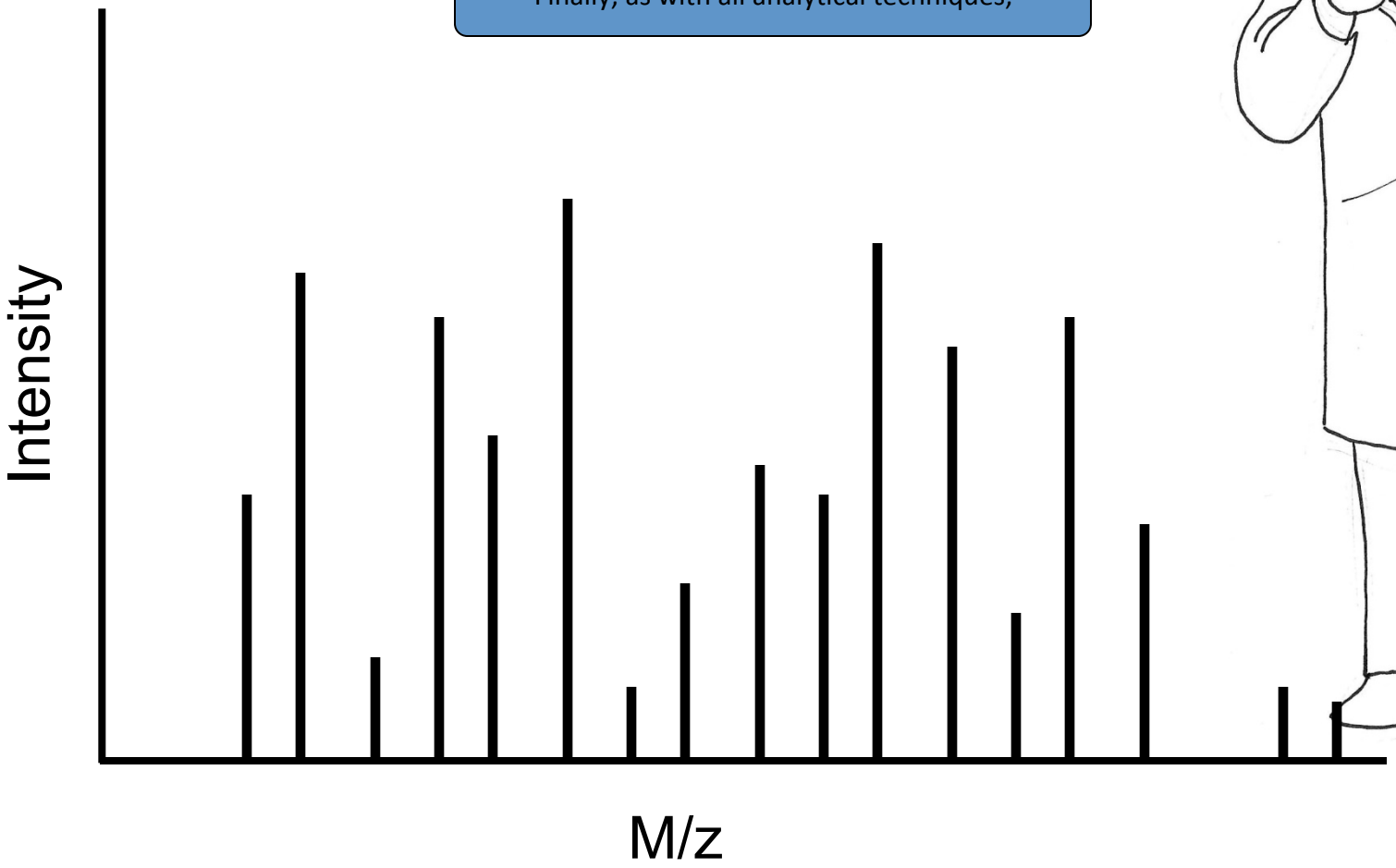


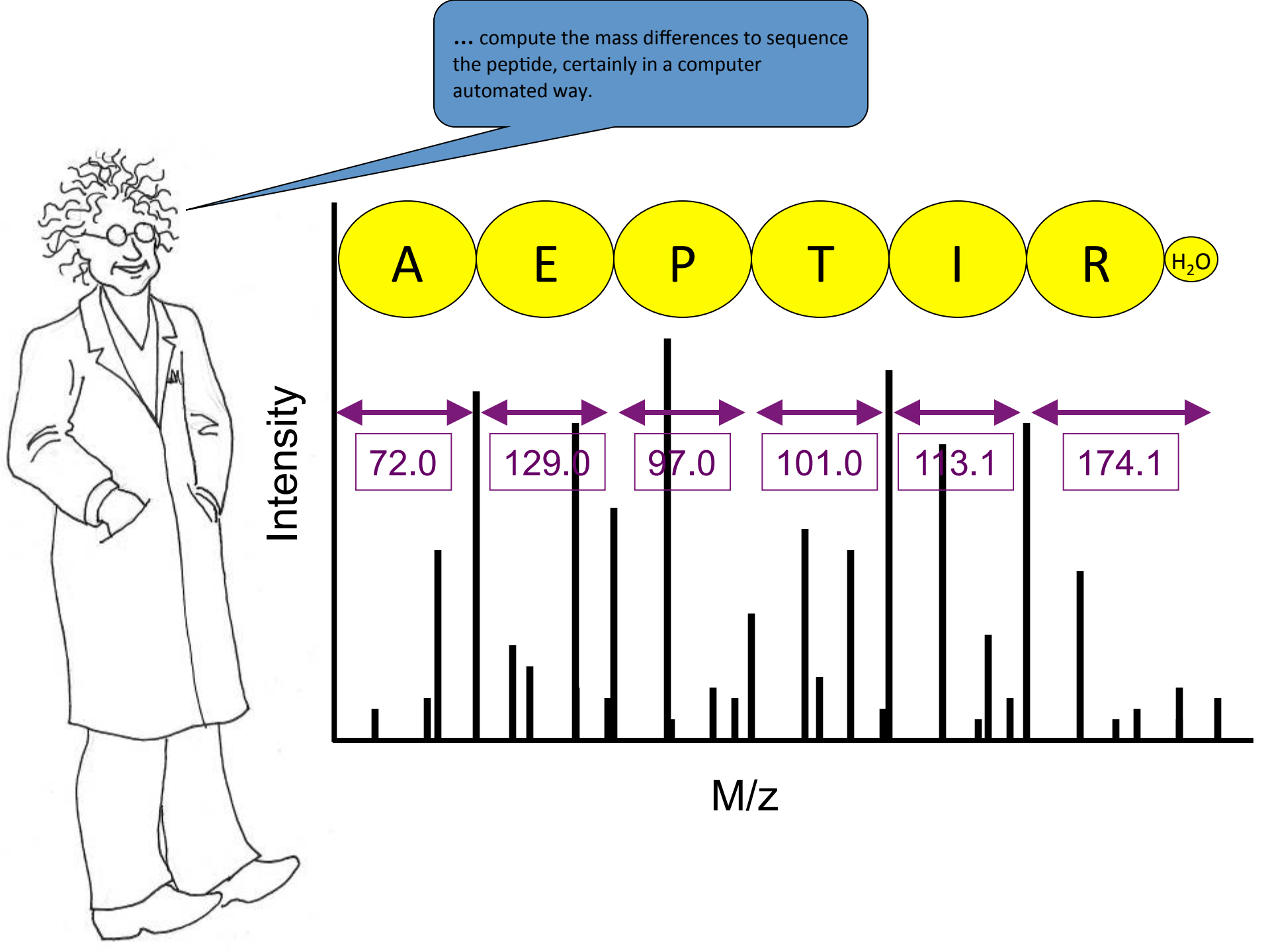
Y-type Ions


... All the peaks have different measured heights and many peaks can often be missing.
H2O
R
I
T
P
E
A
Intensity
M/z

What type of ion they are, making the mass differences approach even more difficult.

Finally, as with all analytical techniques,




… compute the mass differences to sequence the peptide, certainly in a computer automated way.
A
E
P
T
I
R
H_2O
Intensity
72.0
129.0
97.0
101.0
113.1
174.1
M/z

Scaffold 4 MS/MS

- Open Scaffold 4 program
- If asked about database access, cancel
- Run Demo
- Select Tutorial 1
- Click on any protein
- Go to Proteins tab
- Lower panel Click on Spectrum
- Click on Fragmentation Table for theoretical ions
- Change peptides by selecting in upper right pane

Sequencing Explosion

- **1977** Shotgun sequencing invented, bacteriophage *fX174* sequenced.

Eng, J. K.; McCormack, A. L.; Yates, J. R. III
J. Am. Soc. Mass Spectrom. **1994**, 5, 976-989.

- **1989** Yeast Genome project announced
- **1990** Human Genome project announced
- **1992** First chromosome (Yeast) sequenced
- **1995** *H* quenced
- **1996** Yeast Ge uenced
- **2000** Human Genome draft

In 1994 Eng and Yates published a technique to exploit genome sequencing

for use in tandem mass spectrometry.

And the idea was ...

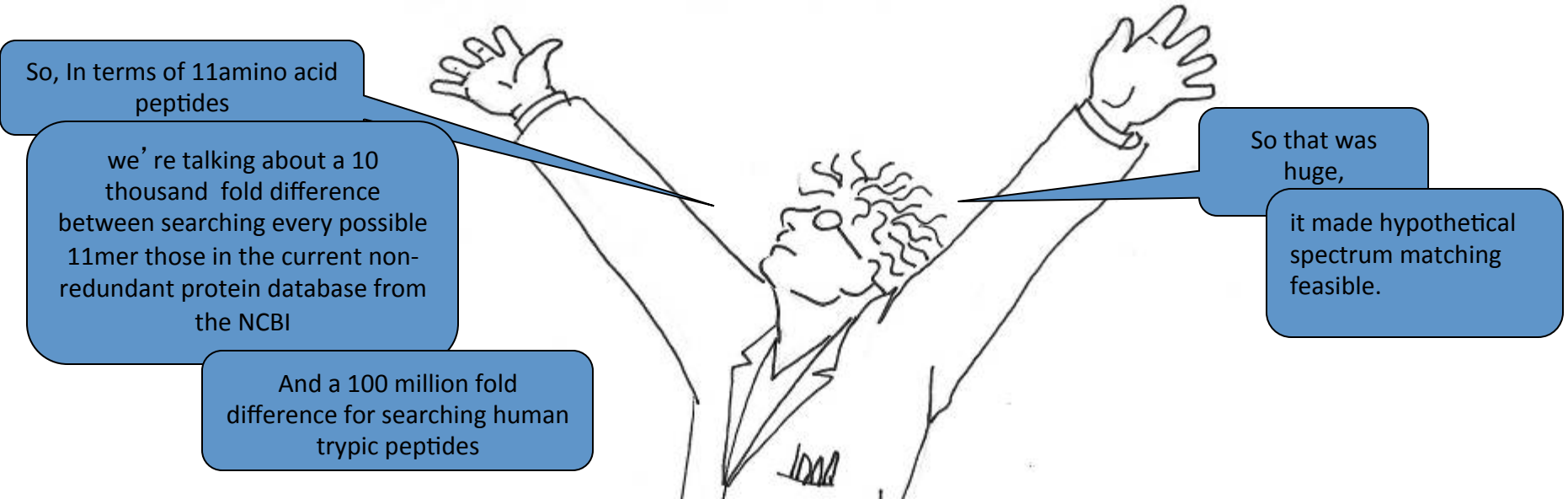
SEQUEST

$2*10^{14}$ -- All possible 11mers (ELVISLIVESK)

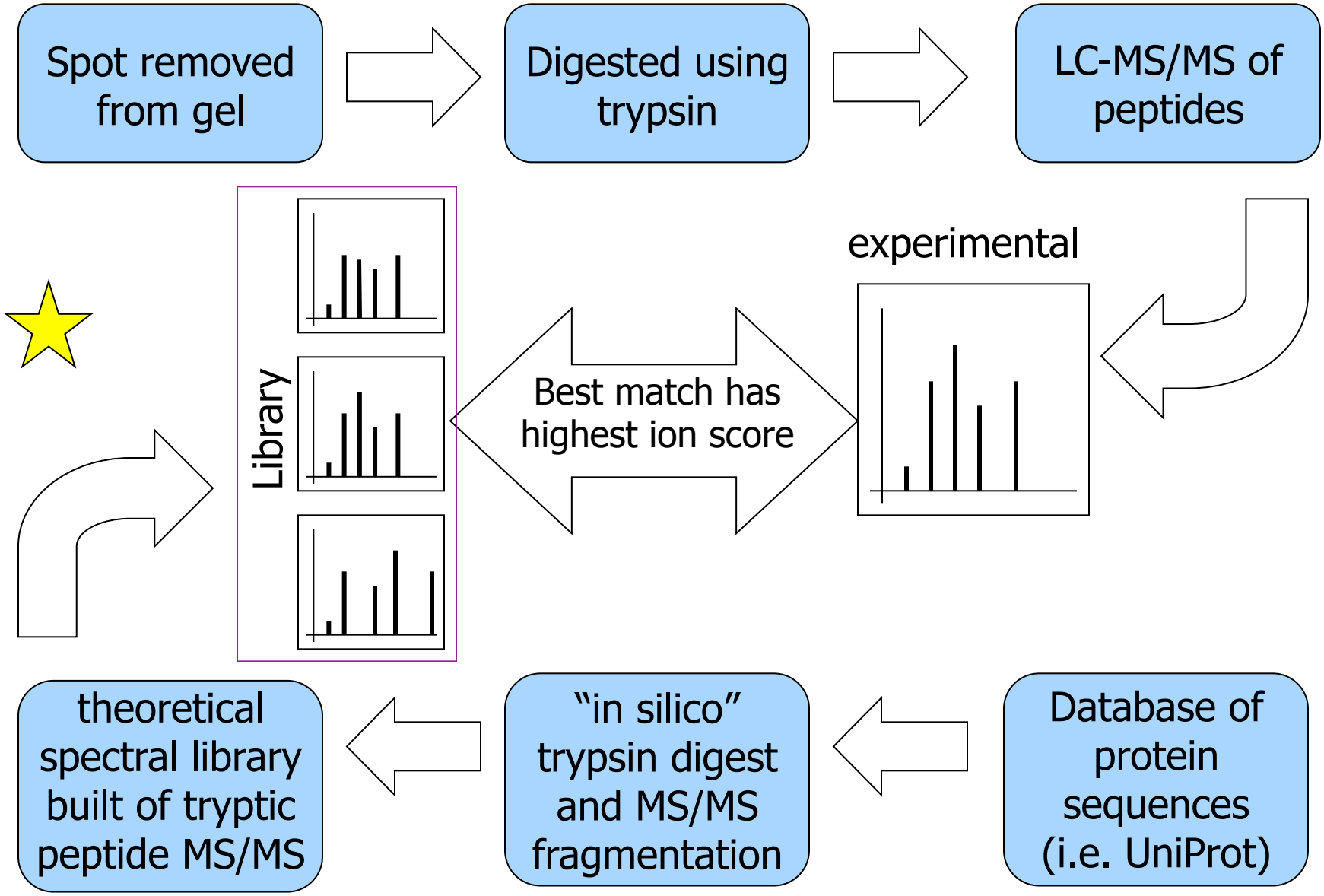
$2*10^{10}$ -- All possible peptides in NR

$1*10^{8}$ -- All tryptic peptides in NR

$4*10^{6}$ -- All Human tryptic peptides in NR



Peptide ID by Spectral Matching Process



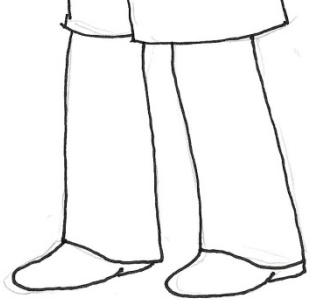
Proteomic Databases:

- UniProt–SwissProt + TrEMBL
- NCBI

MS/MS Search Engines:

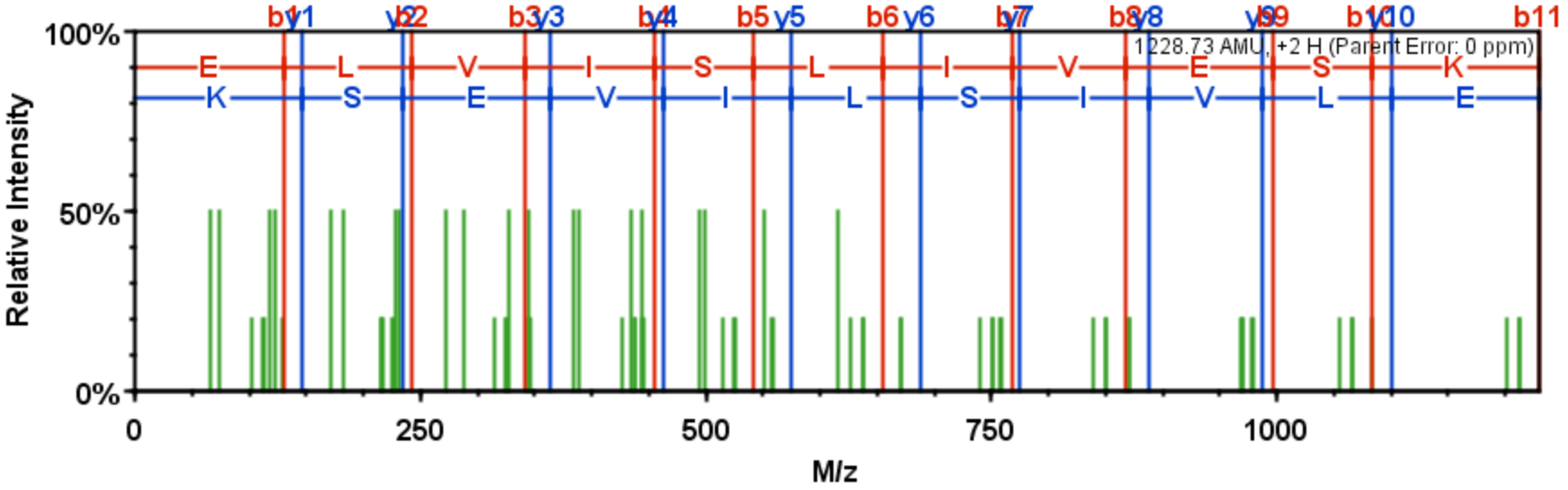
- MASCOT (Matrix Science)
- SEQUEST (J. Eng & J. Yates, Scripps)
- SEQUEST HT (Thermo)
- ProteinProphet (R. Aebersold, ISB)
- OMSSA (NCBI)
- X!Tandem (thegpm)
- MS-Amanda (K. Mechtler, IMP, IMBA & GMI)
- Andromeda (M. Mann, Max Planck Institute)
- Scaffold (Proteome Software) validation only

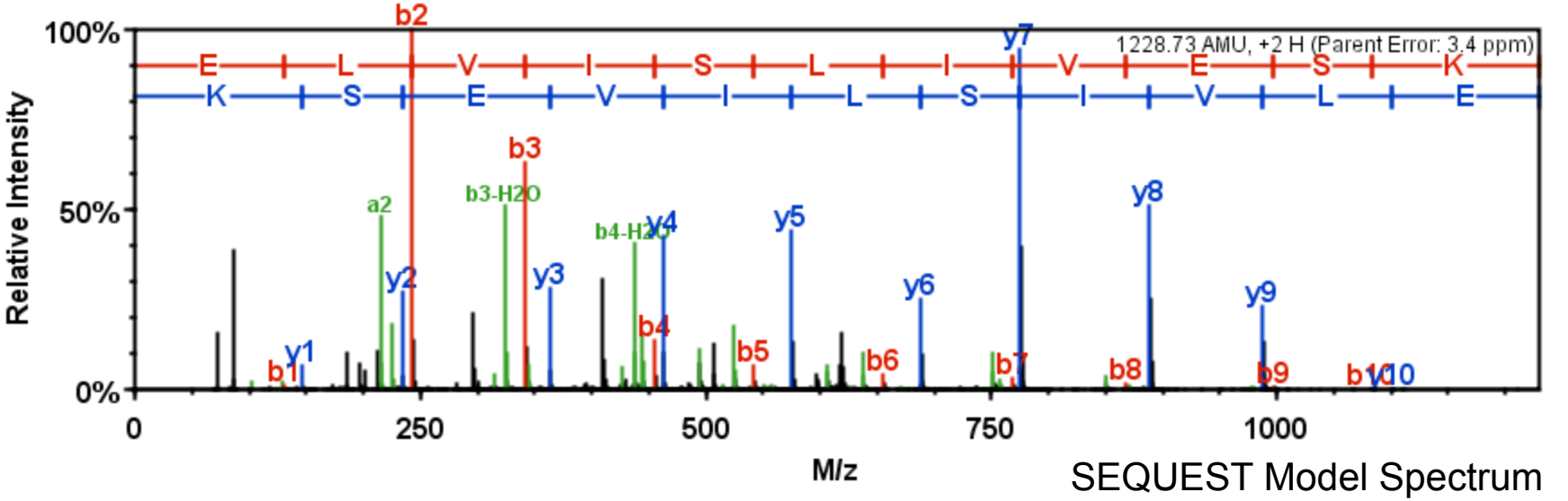
Eng and Yates noted that there was a discontinuity between e intensities of the hypothetical spectrum and the actual spectrum.



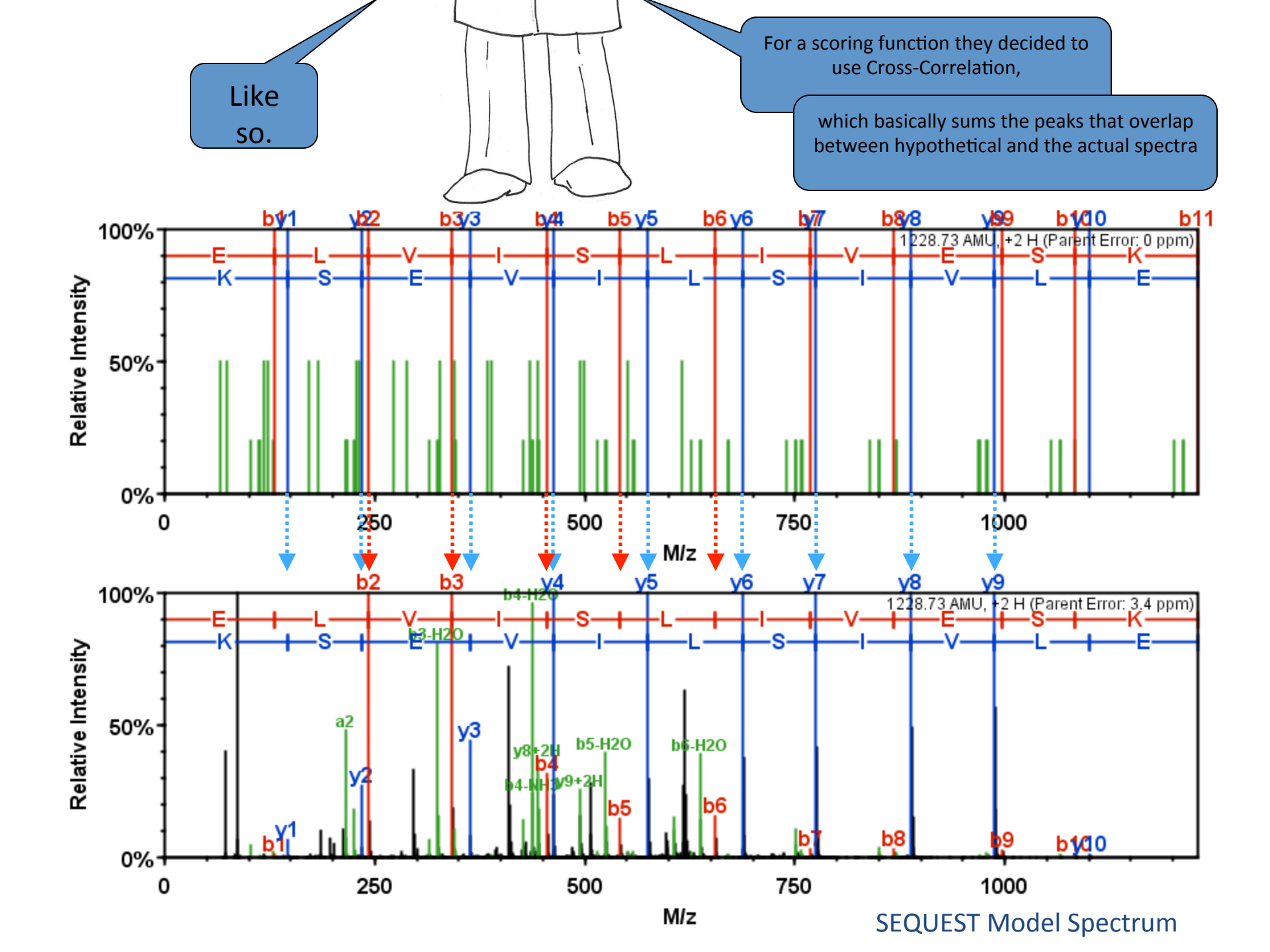
Instead of trying to make a better model,

they decided just to make the actual spectrum look like the model with normalization…



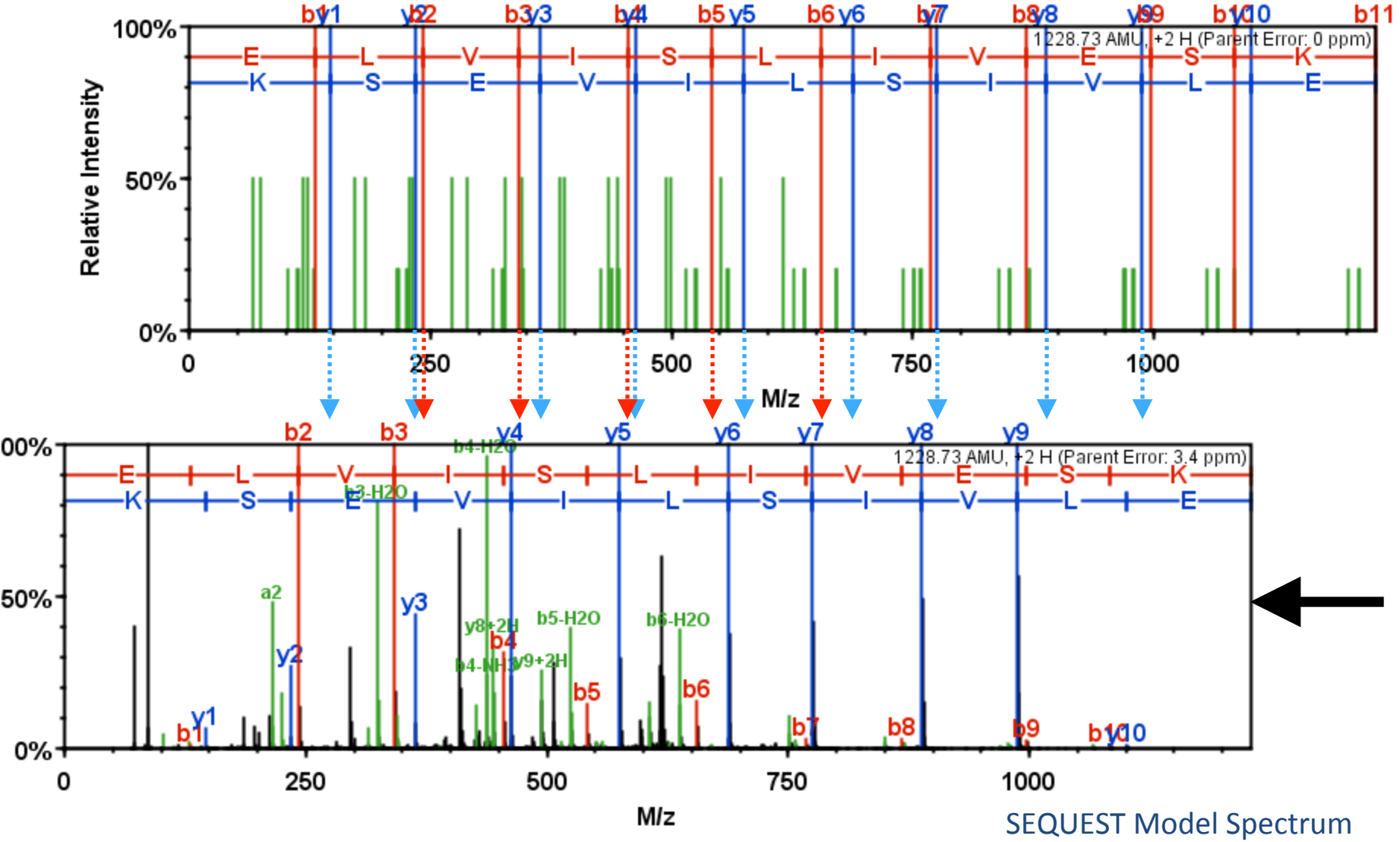


SEQUEST Model Spectrum

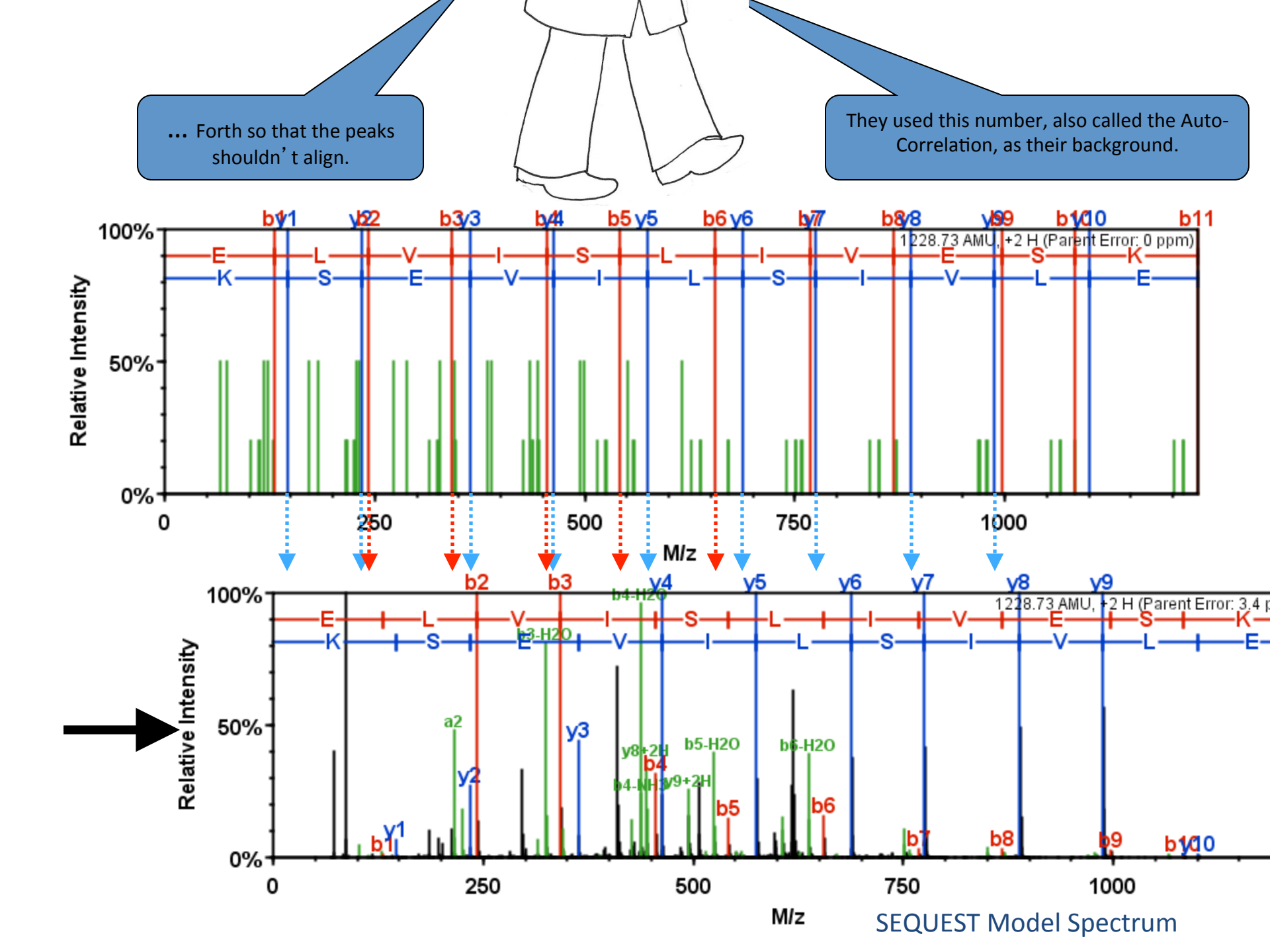


SEQUEST Model Spectrum

And then they shifted the spectra back and ….

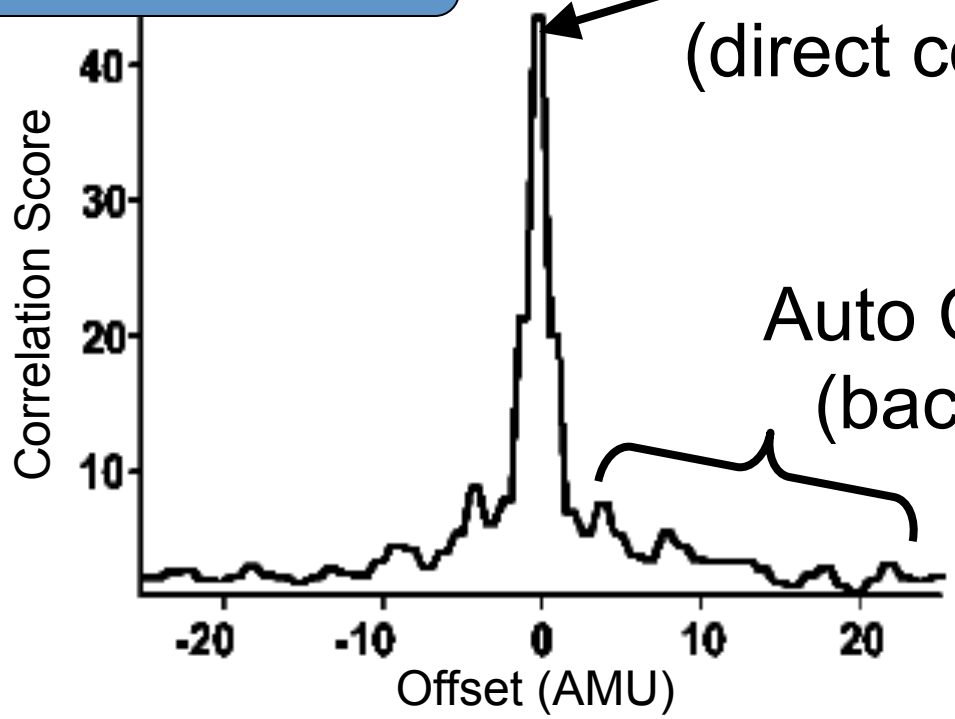


SEQUEST Model Spectrum



SEQUEST Model Spectrum

SEQUEST XCorr



Cross Correlation
(direct comparison)

Auto Correlation
(background)

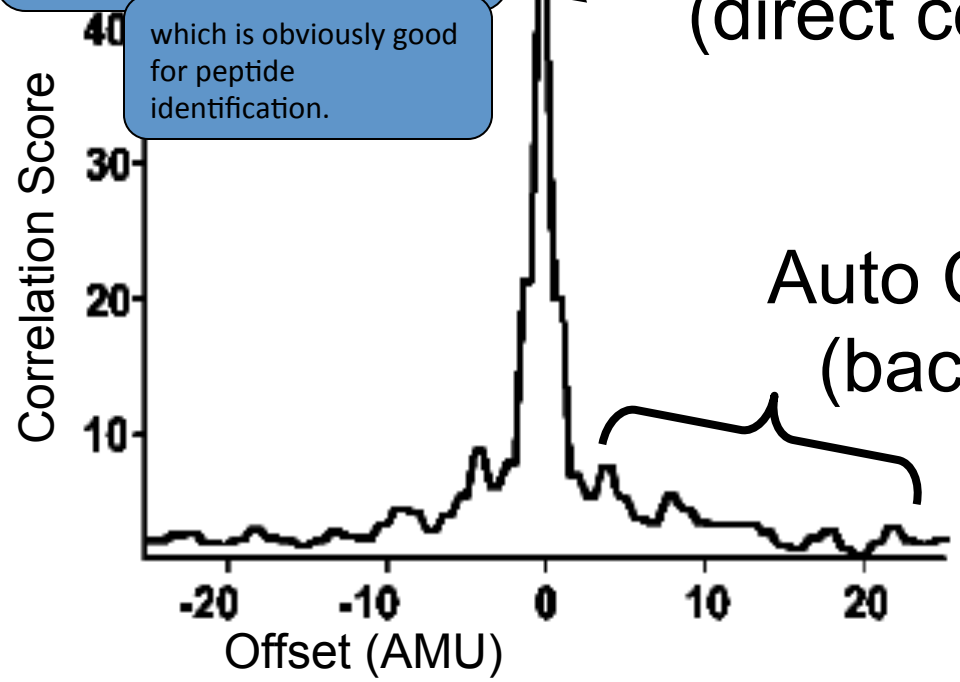
Gentzel M. et al
Proteomics **3** (2003) 1597-1610

SEQUEST XCorr

The XCorr score is the Cross Correlation divided by the average of the auto correlation over a 150 AMU range.

The XCorr is high if the direct comparison is significantly greater than the background,

which is obviously good for peptide identification.



Cross Correlation (direct comparison)

Auto Correlation (background)

$$\text{XCorr} = \frac{CrossCorr}{avg\left(AutoCorr_{\text{offset=-75 to 75}}\right)}$$

Gentzel M. et al
Proteomics 3 (2003) 1597-1610

10 Protein Control Sample (Q-ToF)
Peptide Prophet approach

ALL Other "Best" Matches

Well, Peptide Prophet looks across the entire sample, and not at just one spectrum at a time.

It compares the best match against all of the other best matches in the sample, which is clearly bimodal.

Possibly Correct?

# of Matches

0
100
200
300
400
500
600
700

Mascot: Ion-Identity Score

Keller, A. et al
Anal. Chem. 74, 5383-5392

10 Protein Control Sample (Q-ToF)
Peptide Prophet approach

ALL Other "Best" Matches

The low mode represents matches that are most likely wrong while the high mode represents matches that are probably right.

Possibly Correct?

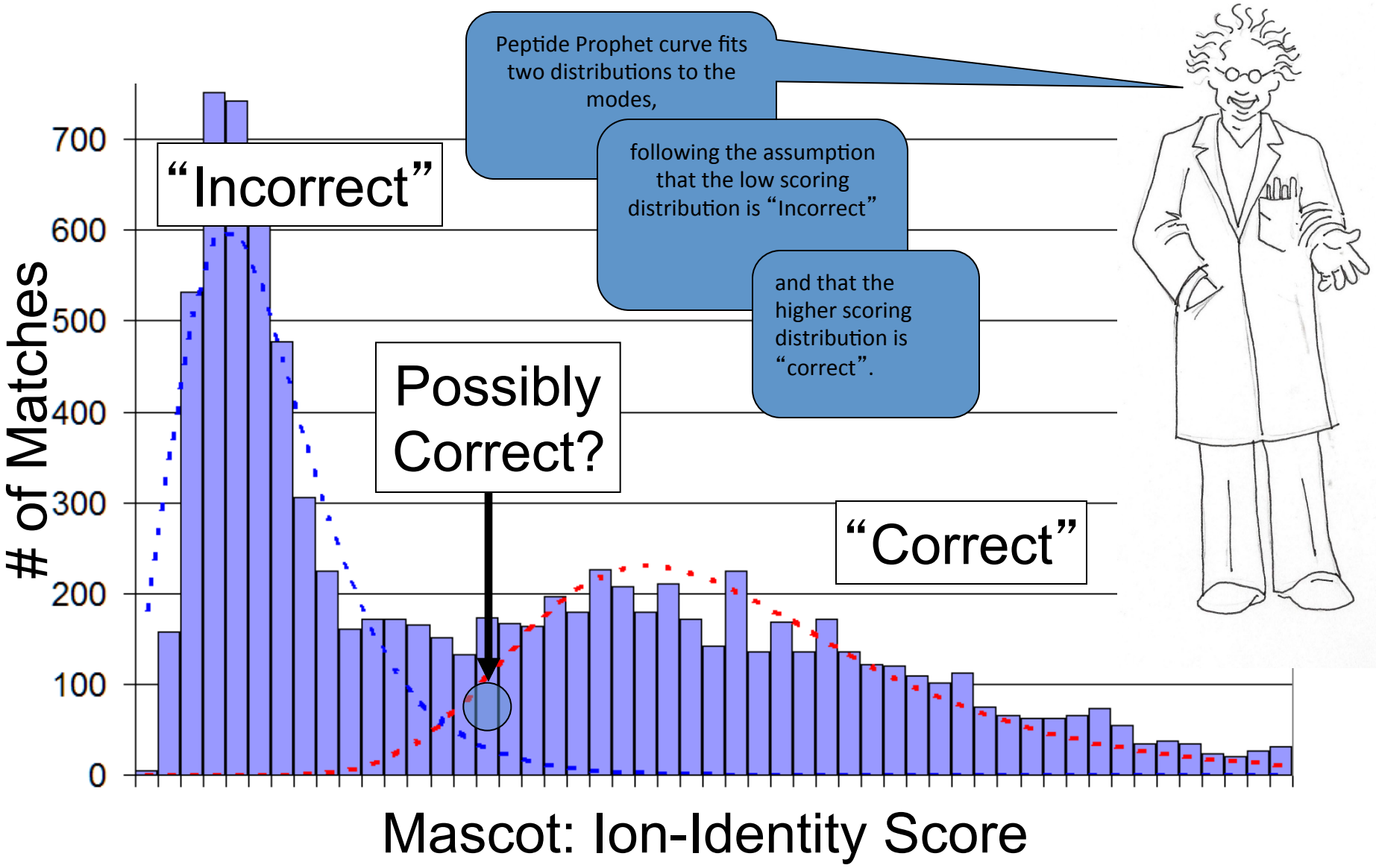
# of Matches

700
600
500
400
300
200
100
0

Mascot: Ion-Identity Score

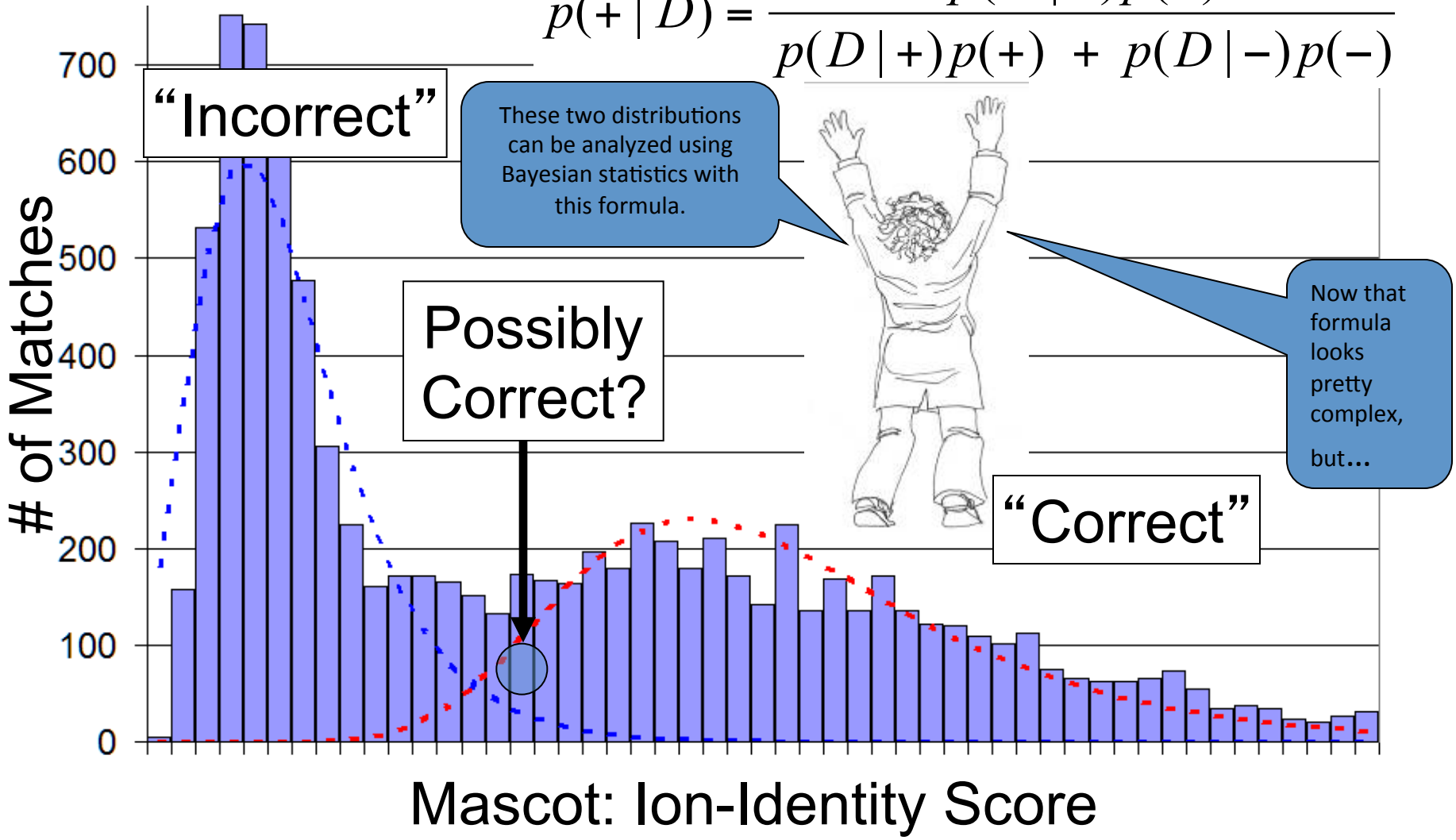
Keller, A. et al
Anal. Chem. 74, 5383-5392

10 Protein Control Sample (Q-ToF)
Peptide Prophet approach

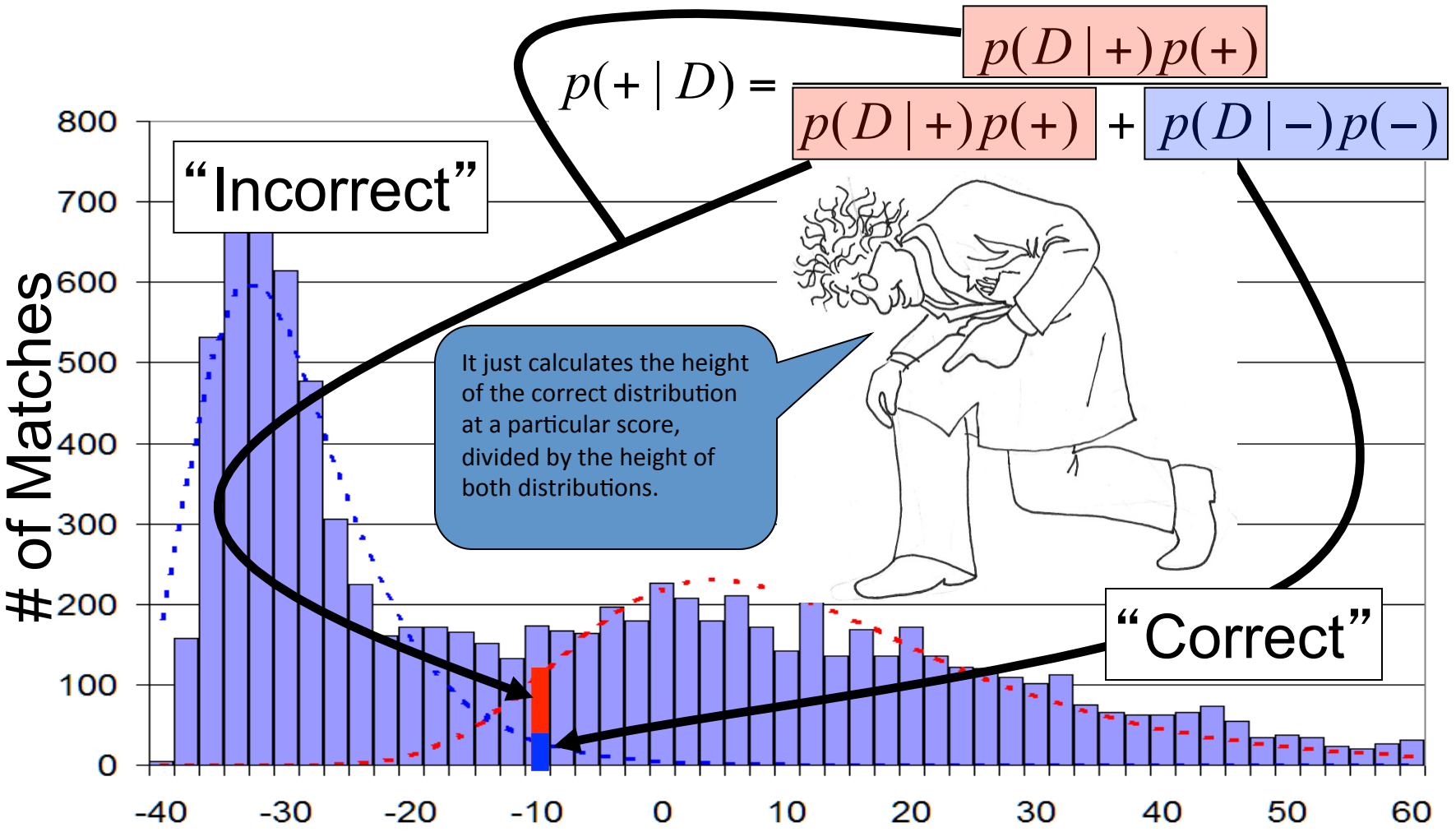


10 Protein Control Sample (Q-ToF)

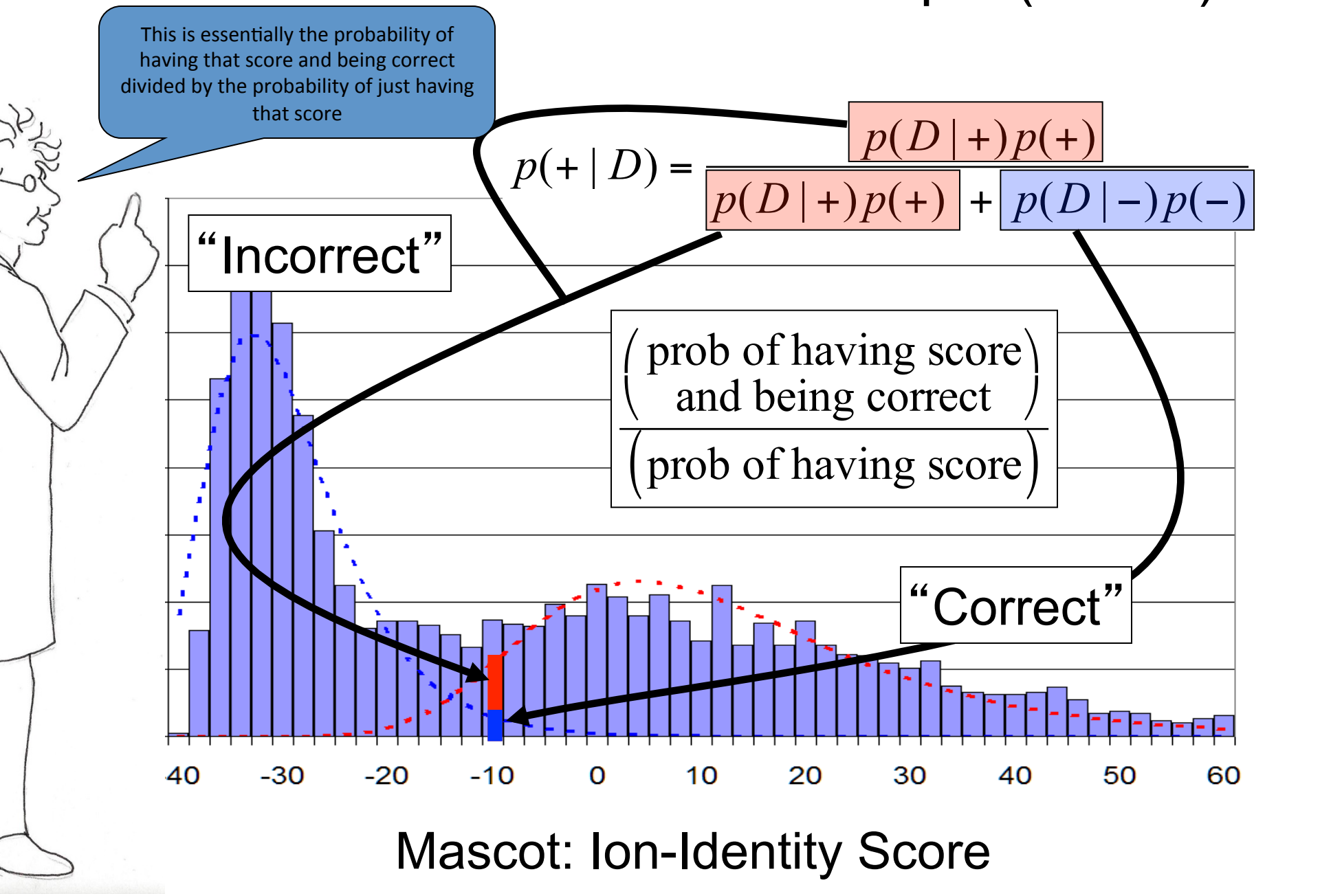
$$p(+\mid D) = \frac{p(D\mid +)p(+)}{p(D\mid +)p(+) \;+\; p(D\mid -)p(-)}$$



10 Protein Control Sample (Q-ToF)



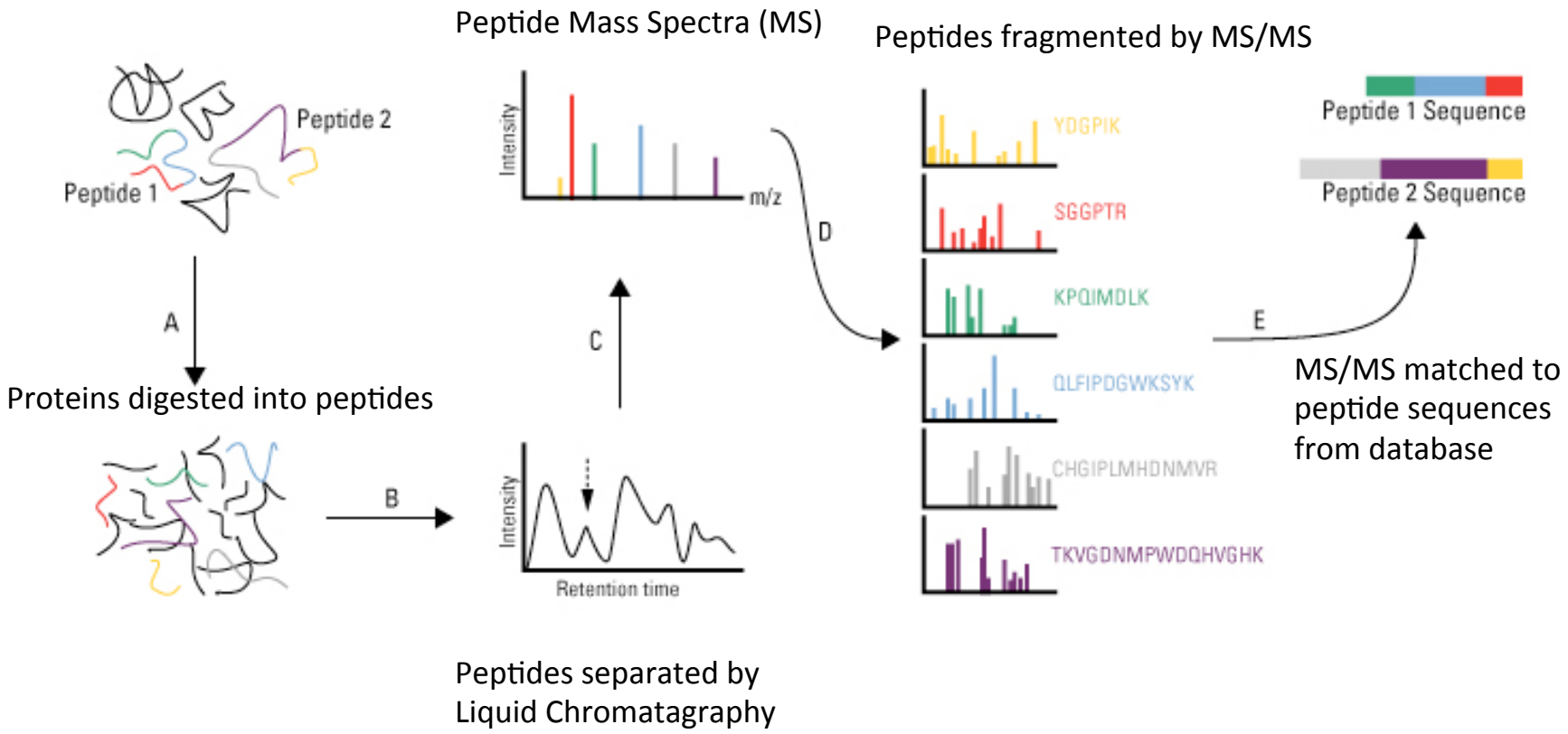
10 Protein Control Sample (Q-ToF)



Scaffold 4 Database Search

- Run Demo, Select Tutorial 1
- Left Pane select Proteins
- Upper right pane shows SEQUEST scores
- Compare Sequest scores and MS/MS
- Left pane Statistics OR Menu Bar Window Statistics
- Lower right pane Prophet distribution of correct and incorrect hits

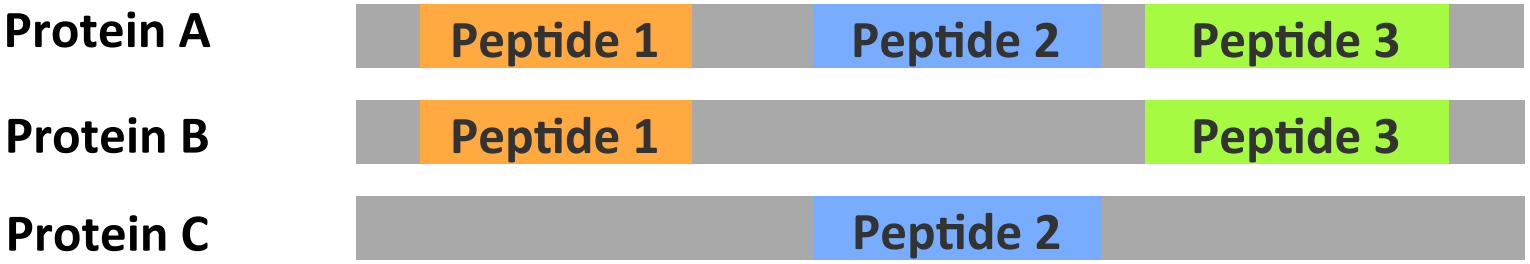
LC-MS/MS Peptide Identification



Protein Inference

General approach is to create a minimal list of proteins.

"Principal of parsimony" or "Occam's razor"



Protein Inference

Peptides identified:

1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISAEK	9	VGINYQPPTVVPGGDLAK
2	AVFVDLEPTVIDEVR	6	AYHEQLSVAEITNACFEPANQMVK	10	AVCMLSNTTAIAEAWAR
3	QLFHPEQLITGKEDAANNYAR	7	YMACCLLYR	11	LDHKFDLMYAK
4	NLDIERPTYTNLNR	8	SIQFVDWCPTGFK		

Assignment of peptides to proteins:

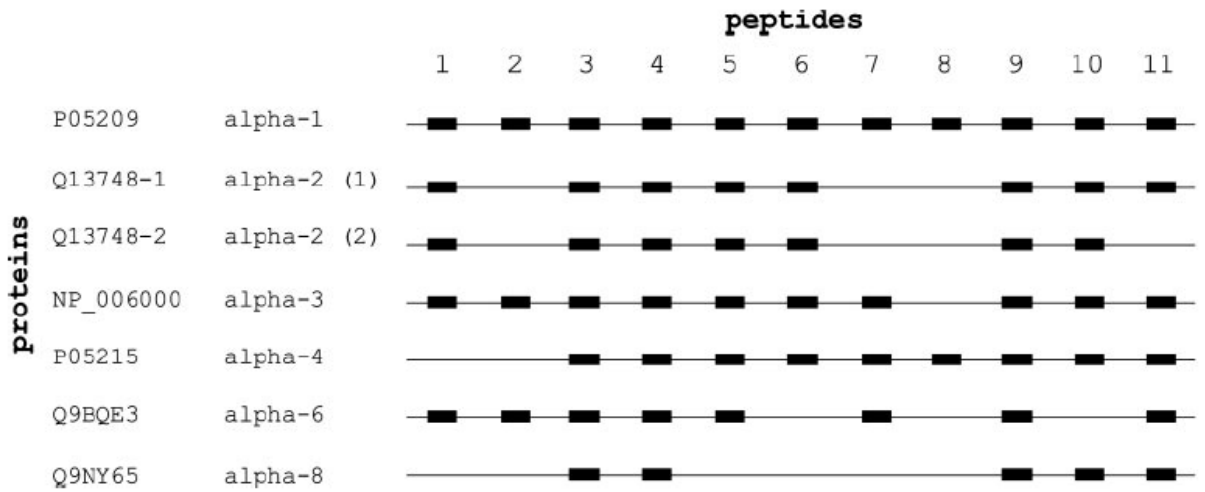
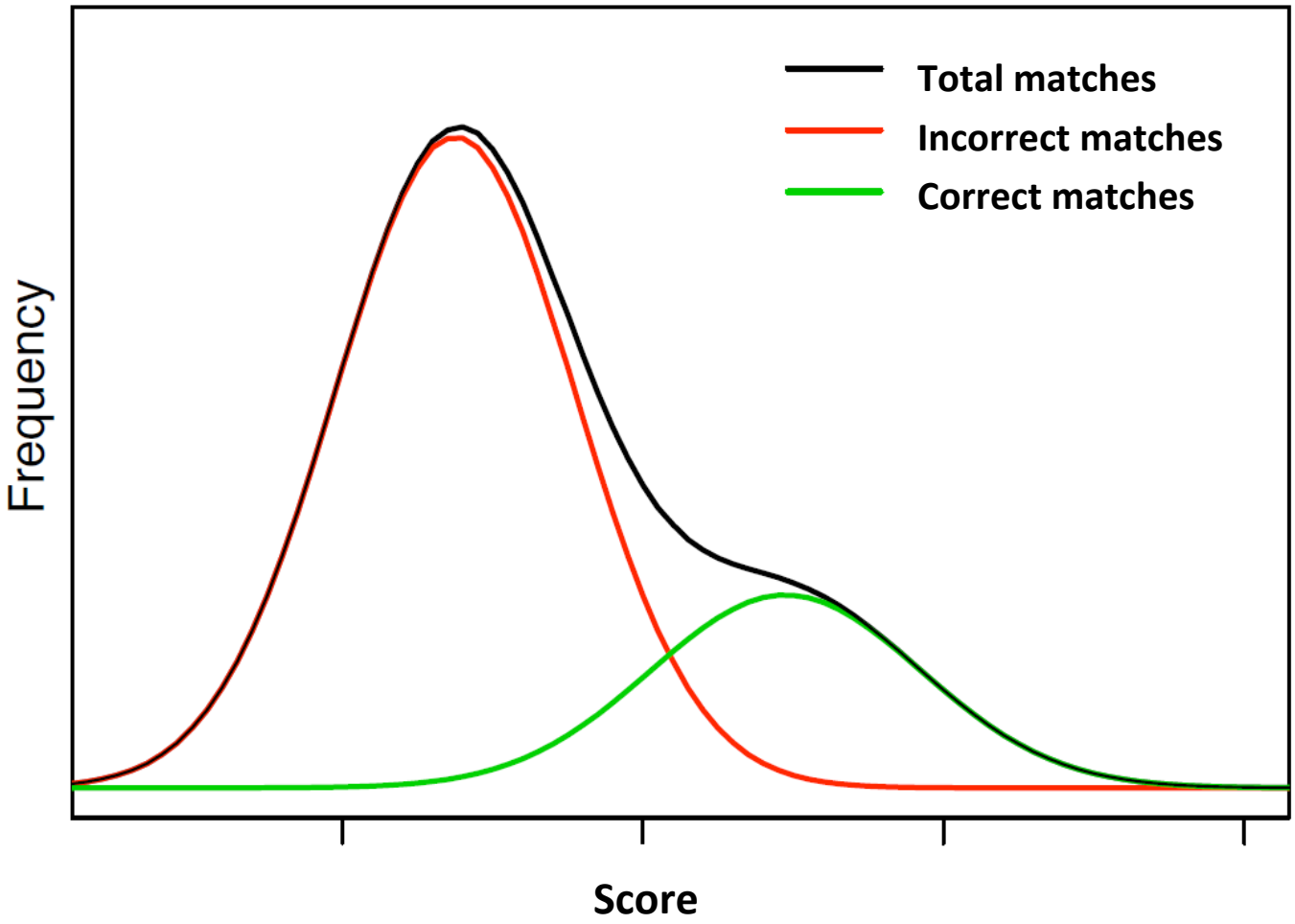


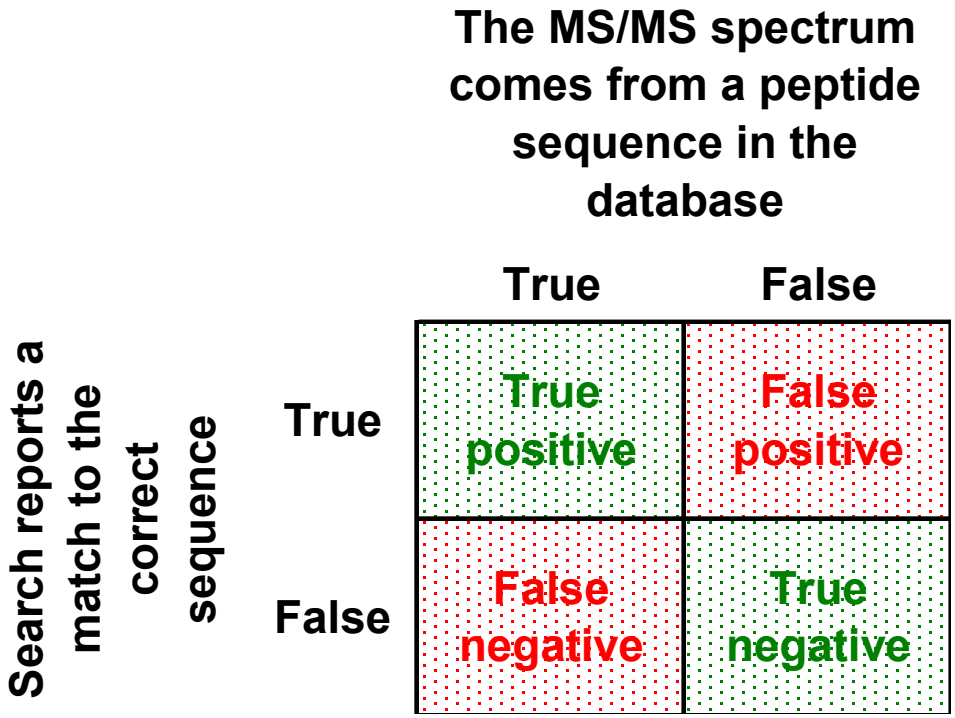
FIG. 3. **An example of a protein family.** Eleven tryptic peptides are identified that are shared between the members of the α-tubulin family. None of the proteins is identified by a peptide that is unique to it, thus making it impossible to determine which particular member(s) of the family is present in the sample.

- **Nesvizhskii, A. I. and Aebersold, R. (2005). Interpretation of shotgun proteomic data - The protein inference problem. Mol. & Cellular Proteomics, 4, 1419-1440.**

Distribution of search engine matches between MS/MS spectra and peptide sequences using true and decoy databases



False Discovery Rate calculated by searching the data with a decoy DB to provide statistical confidence measure for peptide identifications

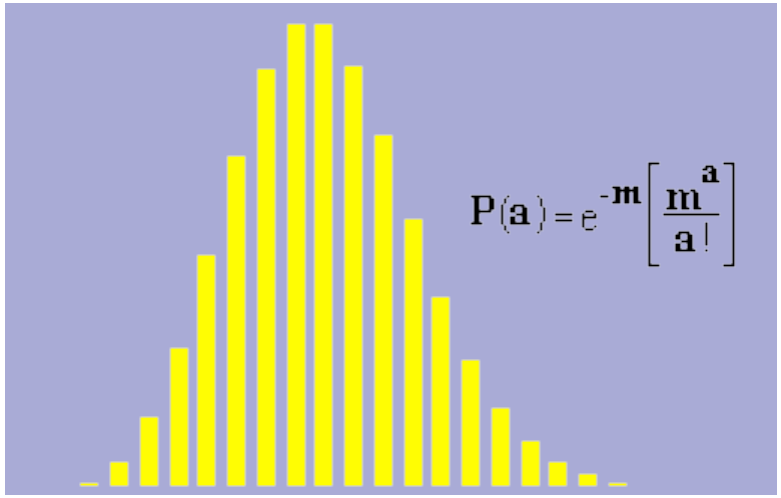


False Discovery Rate
= FP / (FP + TP)

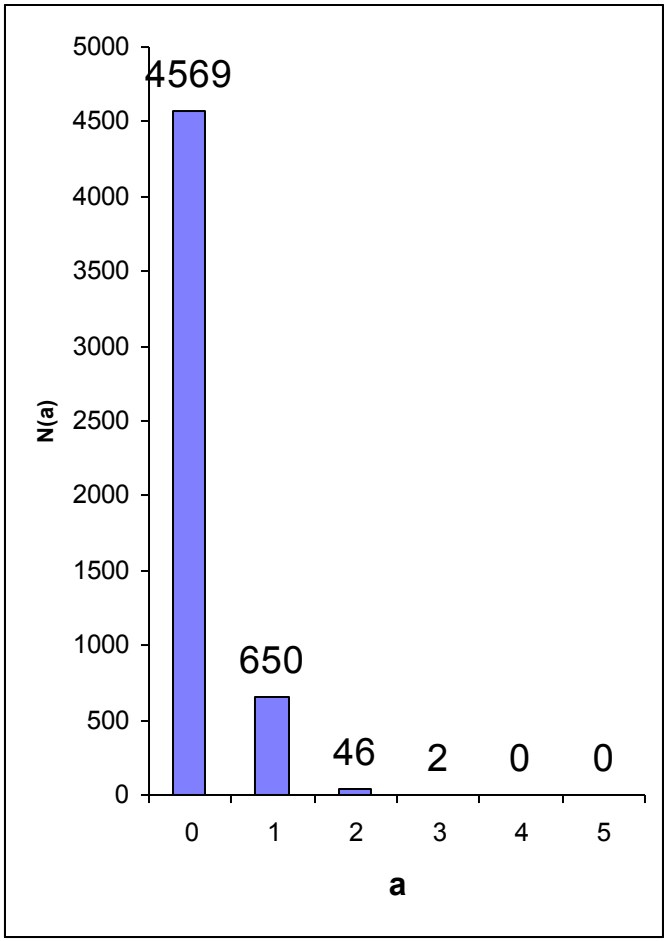
True Positive Rate
= TP / (TP + FN)

False Positive Rate
= FP / (FP + TN)

One Hit Wonders



- Huge MudPIT data set
- Search Swiss-Prot using drosophila taxonomy filter (5268 entries)
- 75,000 matches with 1% FDR
- i.e. 750 false matches

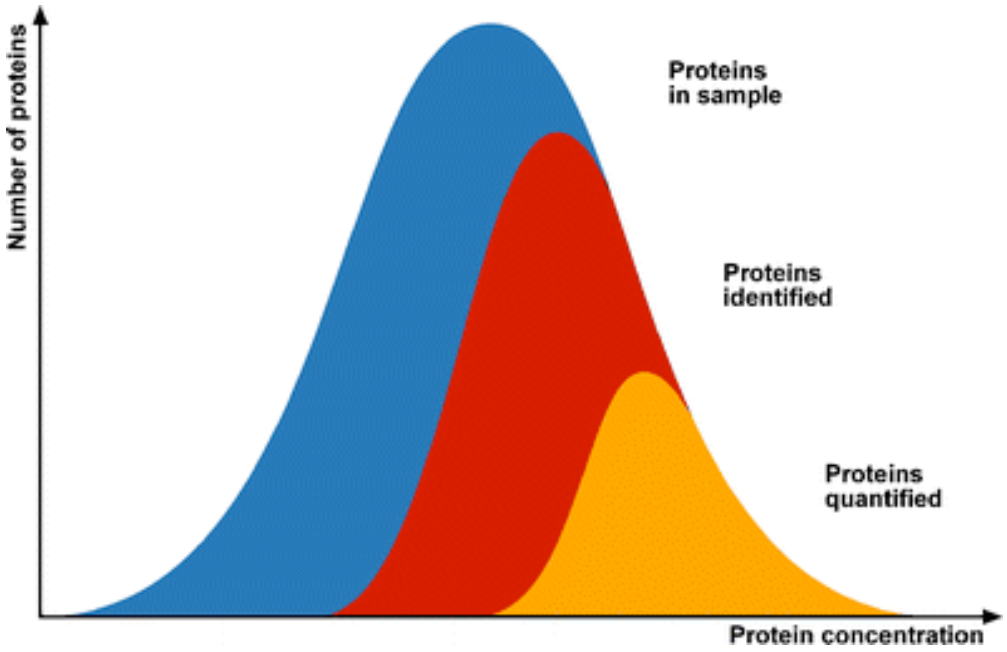


Scaffold 4 Protein Inference and FDR

- Run Demo Select Label-Free
- Samples
- Similarity View
- Change Protein and Peptide Thresholds and Minimum Number of Peptides
- View Pink Box lower left changes in FDR and number of identifications
- Scroll down to find decoy hits if FDR >0

Quantitative Proteomics

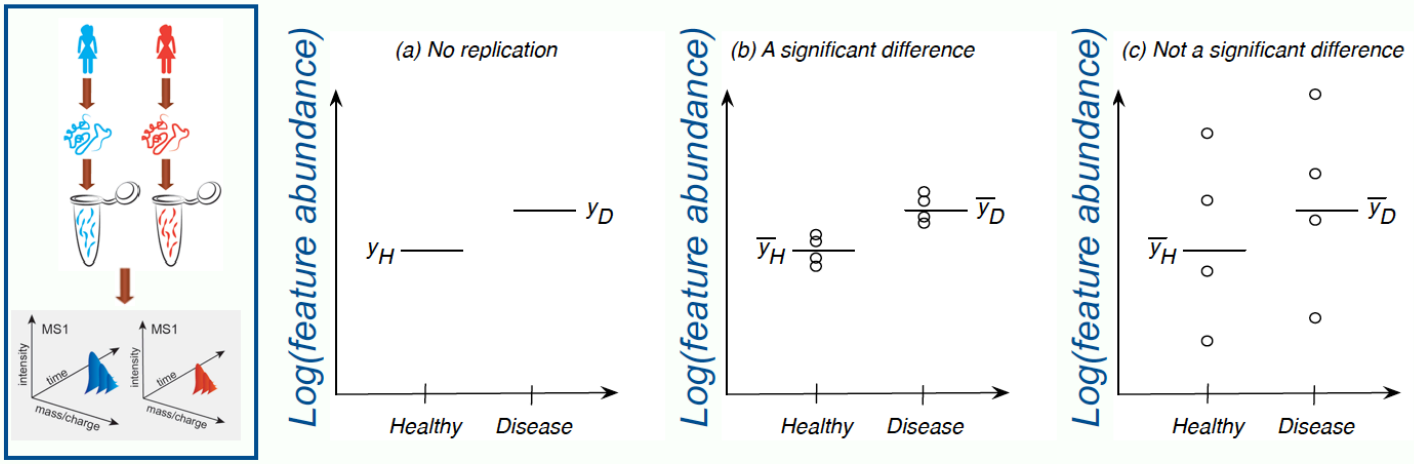
Quantifiable Proteins Are Subset of Proteome



- Spectral Counting—relatively quick and inexpensive, excellent choice for pilot experiment requiring no special sample prep
- TMT/iTRAQ labeling—good precision, minimize sampling differences by combining samples into one LC-MS/MS run

PRINCIPLE 1: REPLICATION

(1) carries out the inference and (2) minimizes inefficiencies



Two levels of randomness imply two types of replication:

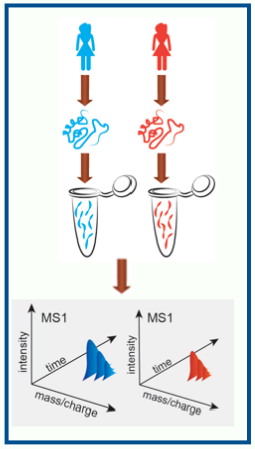
- *Biological replicates:* selecting multiple subjects from the population
- *Technical replicates:* multiple runs per subject

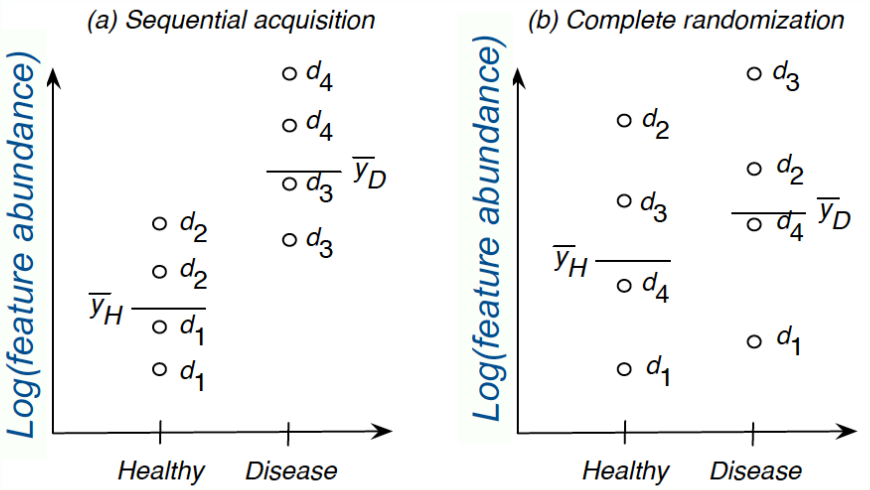
Oberg and Vitek, *J. Proteome Research*, 8, 2009

From Olga Vitek, US HUPO 2016, "Statistics for (Targeted, Label-Free) Proteomics"

PRINCIPLE 2: RANDOMIZATION

Prevents bias





No randomization
= confounding
= bias

Complete randomization
= no bias

Two levels of randomness imply two types of randomization:

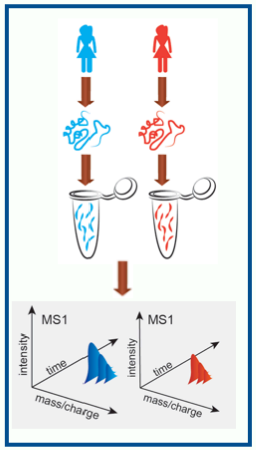
- *Biological replicates:* random selection of subjects from the population
- *Technical replicates:* random allocation of samples to all processing steps

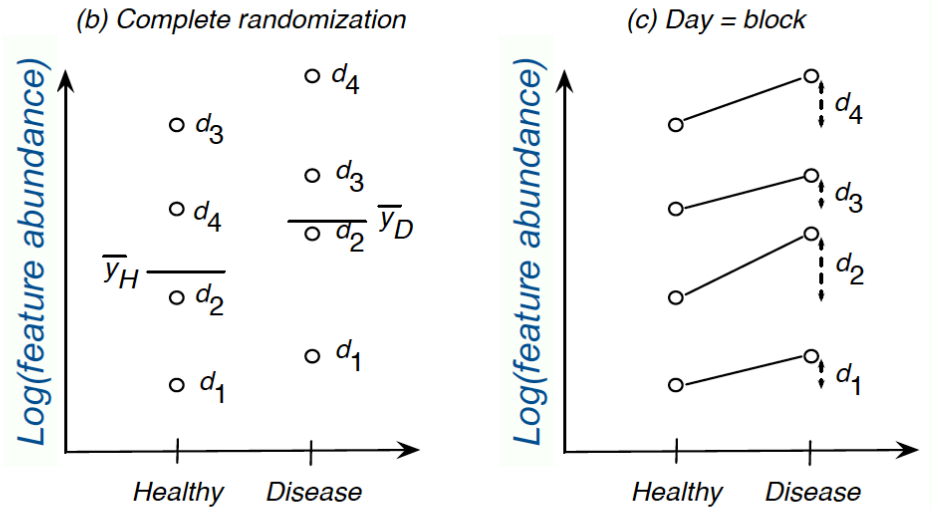
Oberg and Vitek, *J. Proteome Research*, 8, 2009

From Olga Vitek, US HUPO 2016, "Statistics for (Targeted, Label-Free) Proteomics"

PRINCIPLE 3: BLOCKING

Helps reduce both bias and inefficiency





Complete randomization
= inflated variance

Block-randomization
= restriction on randomization
= systematic allocation

Two levels of randomness imply two types of blocks:

- *Biological replicates:* subjects having similar characteristics (e.g. age)
- *Technical replicates:* samples processed together (e.g. in a same day)

Oberg and Vitek, *J. Proteome Research*, 8, 2009

From Olga Vitek, US HUPO 2016, "Statistics for (Targeted, Label-Free) Proteomics"

FINAL THOUGHT: SIMPLICITY IS GOOD

Complicated methods fail for complicated reasons

- A combination of a complicated algorithm and small sample size
- Problems hard to detect
  - The paper eventually retracted

http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

nature medicine

Nature Medicine **12**, 1294 - 1300 (2006)
Published online: 22 October 2006 | Corrected online: 27 October 2006 | Corrected online: 21 July 2008 | Retracted: 07 January 2011 | doi:10.1038/nm1491

There is a Corrigendum (November 2007) associated with this Article.

There is a Corrigendum (August 2008) associated with this Article.

There is a Retraction (January 2011) associated with this Article.

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1,2,3], Johnathan Lancaster[4] & Joseph R Nevins[1,2,3]

ARTICLE LINKS
- Supplementary info

ARTICLE TOOLS
- Send to a friend
- Export citation

From Olga Vitek, US HUPO 2016, "Statistics for (Targeted, Label-Free) Proteomics"

Label Free Spectral Counting

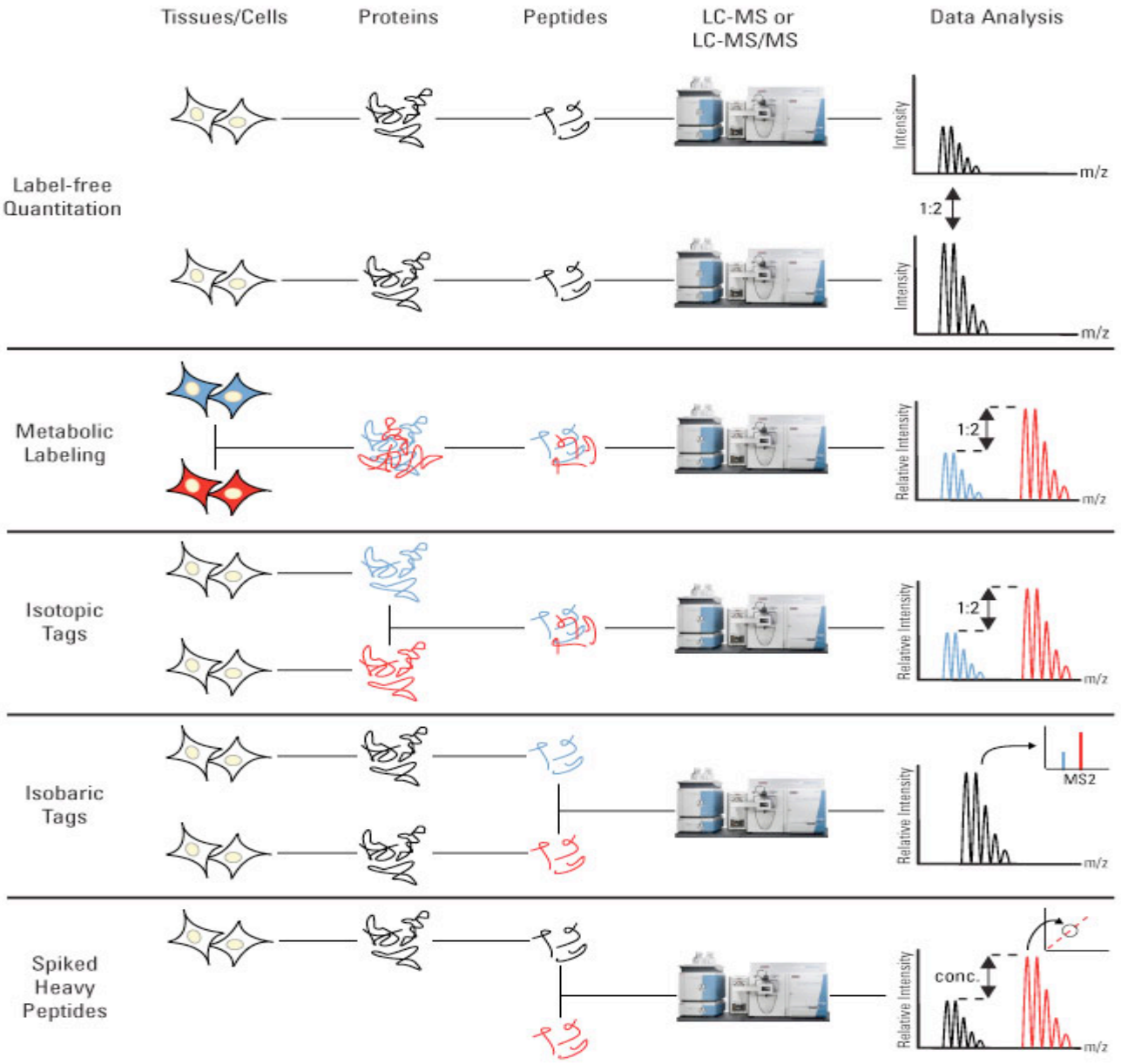
- Relative quantitation: Normalized PSMs used to compare samples

- Absolute quantitation: Approximate effect of protein length using APEX, NSAF, emPAI
- emPAI PAI= $N_{observed}/N_{observable}$
- emPAI = $10^{PAI} - 1$

Scaffold 4 Spectral Counting

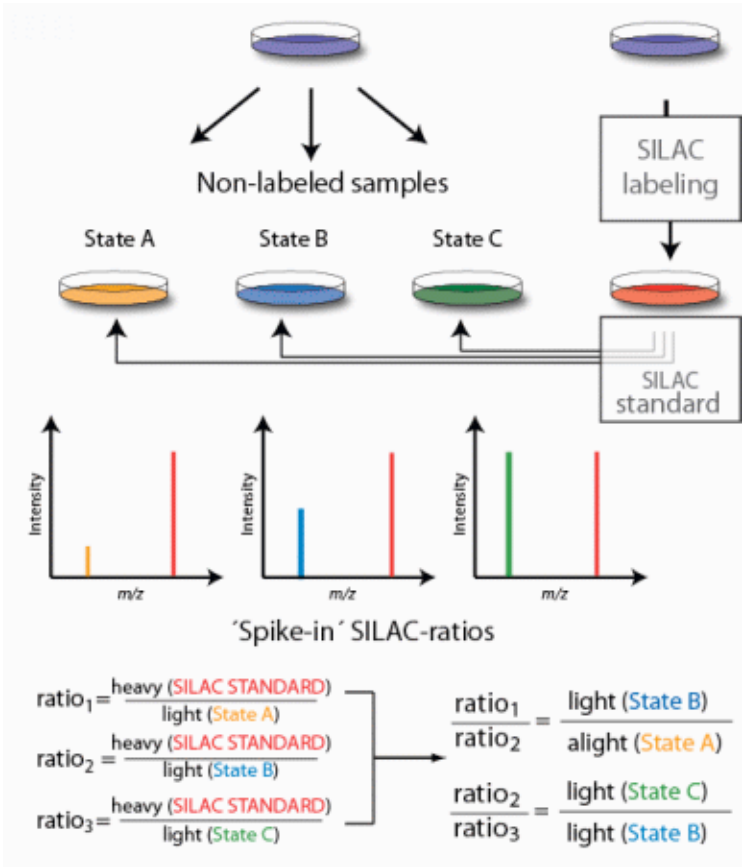
- Run Demo Select Label-Free
- Left pane Samples
- View Menu Uncheck Show GO Annotations
- Display Options Total Spectral Count
- Note proteins 4 and 5 have similar counts
- Experiment Quantitative Analysis
- Quantitative Method emPAI
- Select Compare Categories
- Fisher's exact test

Quantitation Methods

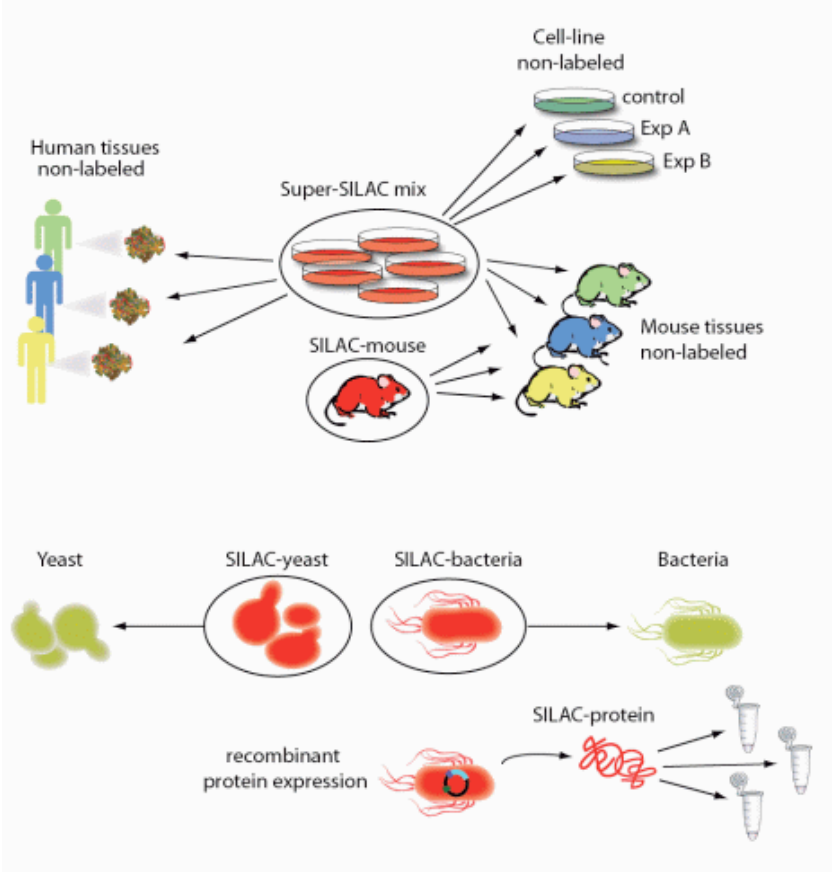


http://www.piercenet.com/method/quantitative-proteomics

Spike-in SILAC standard

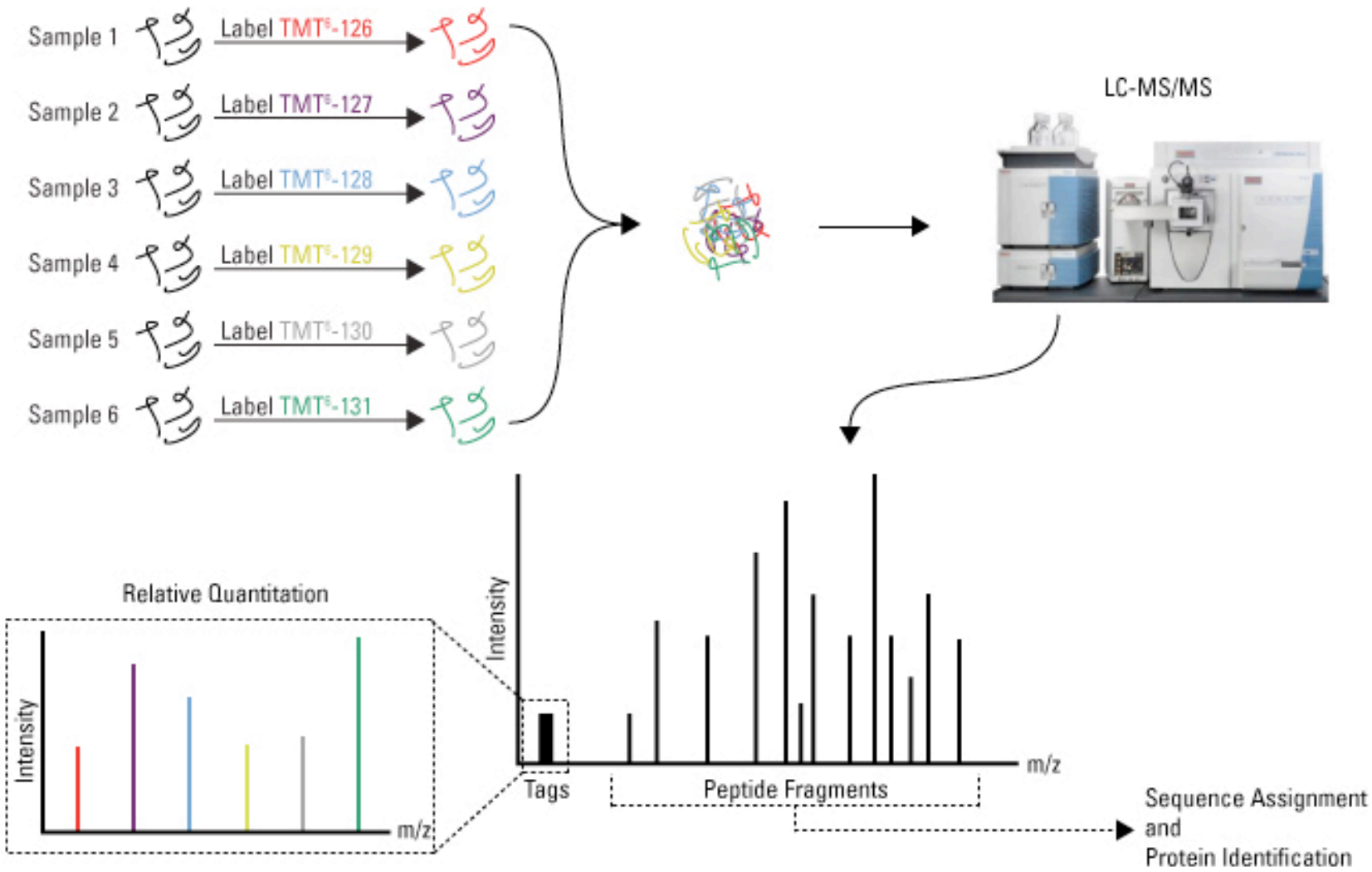


Super SILAC for Tissue quantitation



Matthias Mann, Max Planck Institute for Biochemistry
http://www.biochem.mpg.de/mann/SILAC/index.html

Isobaric Tagging: iTRAQ/TMT

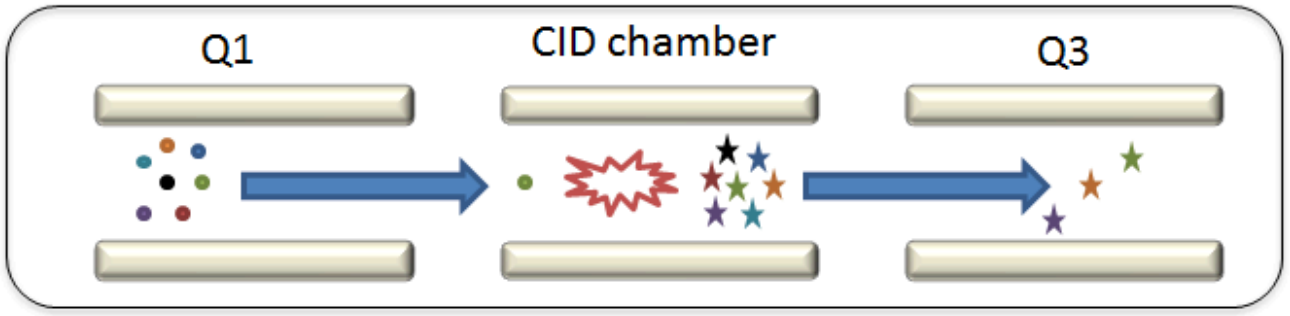


http://www.piercenet.com/method/quantitative-proteomics

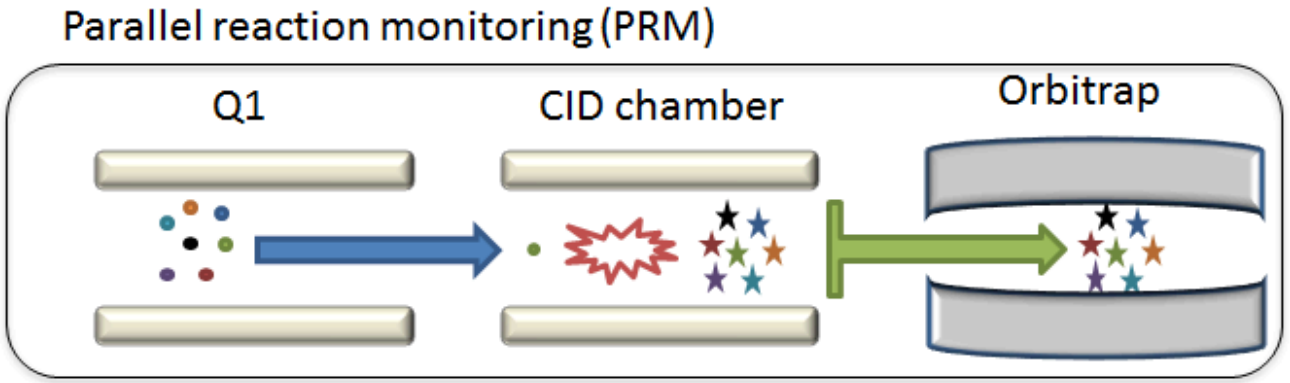
Scaffold 4 iTRAQ

- Run Demo iTRAQ
- Q+ menu button opens new program Q+
- Select #1 protein Apolipophorins
- Switch to Proteins on left sidebar
- Upper pane shows peptides
- Lower pane shows quantitation
- Lower pane select Spectrum and compare peptides

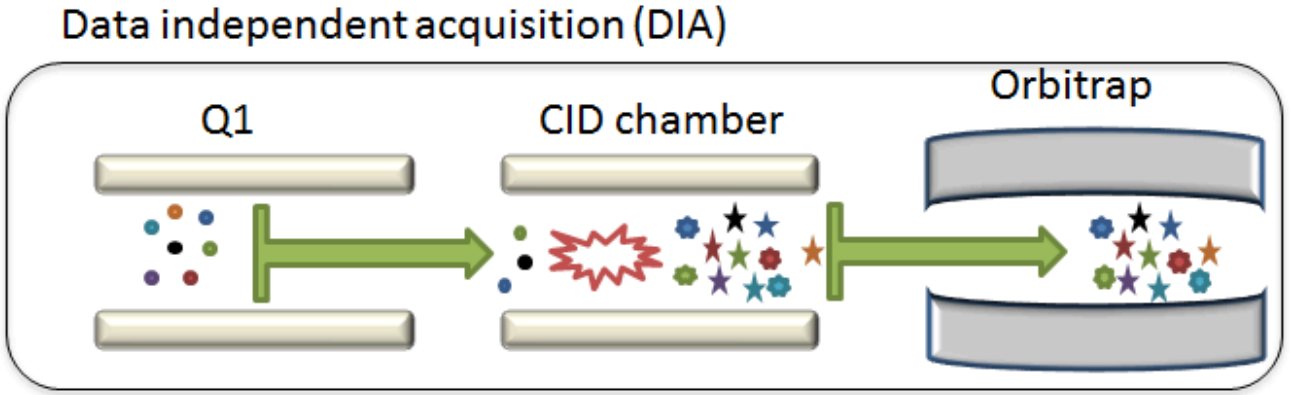
QQQ
1 precursor ion
1-5 product ions



Orbitrap
1 precursor ion
All product ions

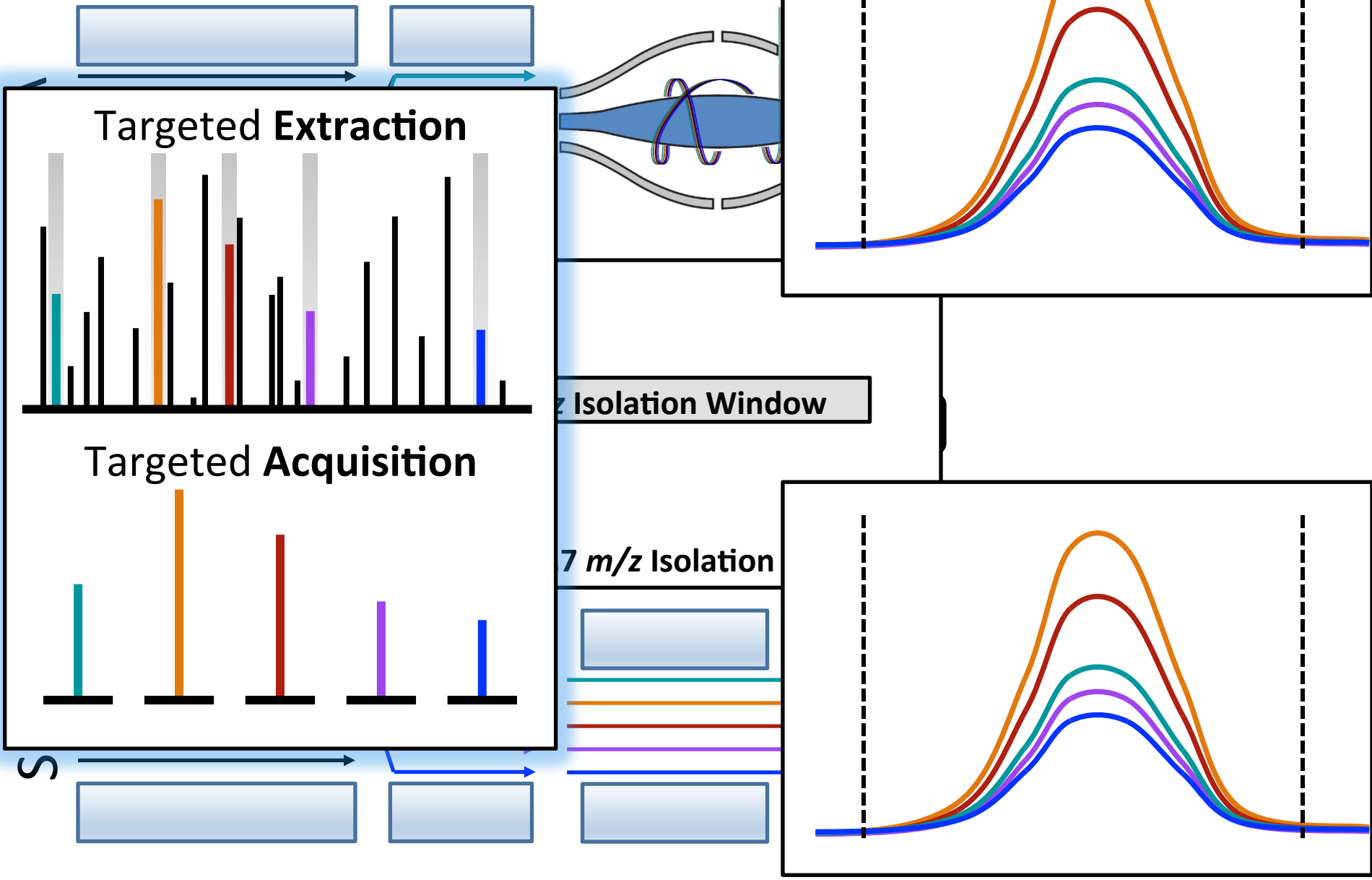


Orbitrap
5-20 m/z
precursor window
All product ions



E. Nouri-Nigjeh et al., Chapter 5: Targeted Proteomics in Translational and Clinical Studies, in Biochemistry, Genetics and Molecular Biology » "Recent Advances in Proteomics Research"

DIA compared to SRM/PRM

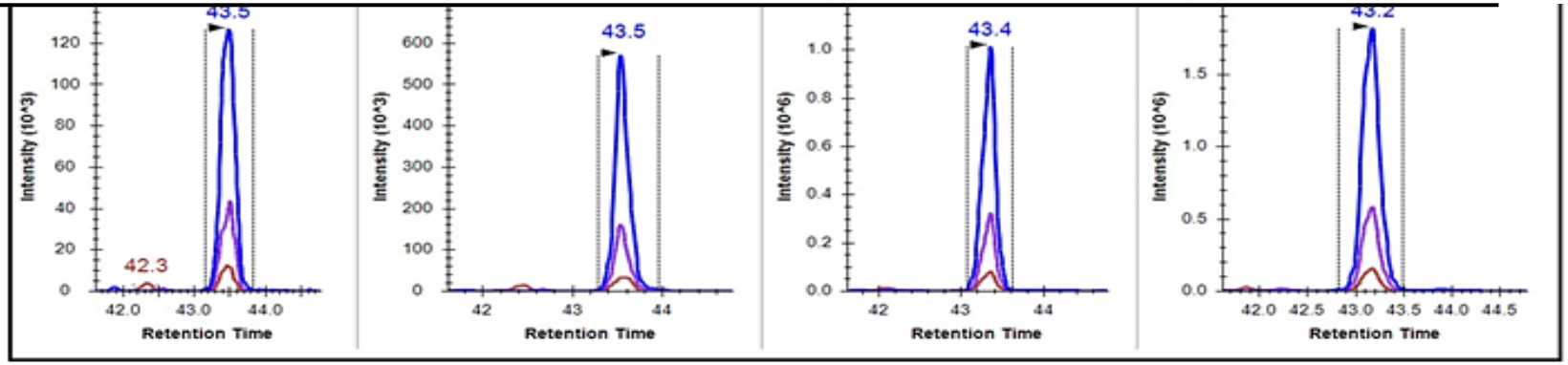


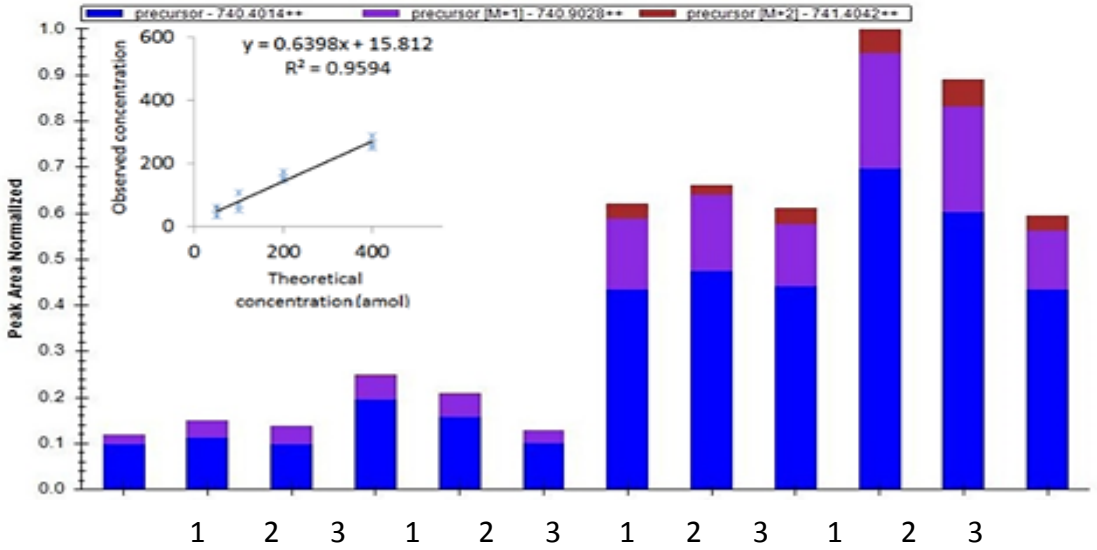
Jarrett Egertson, from Skyline page Slides Explaining Data Independent Acquisition

Targeted Quant using PRM with Skyline

BSA

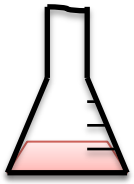
50 amol | 100 amol | 200 amol | 400 amol



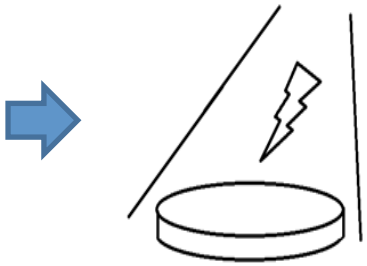


Replicate peak areas

Lydia Contreras Quantitation Workflow

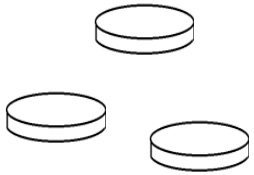


D. radiodurans were cultured to exponential (OD = 1) or stationary (OD = 3) phase in 30 degree shaker.

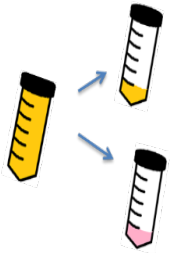


Cells were kept cold on ice and irradiated under 0, 2, 5 & 15 kGy (250Gy/s) with a 10 MeV, 18 kW LINAC β ray source.

Cells were diluted 4-5 fold to OD ~1 and recovered in fresh culture (TGY) medium for 2 hours at 30° C.



Cells were plated on TGY plates and incubated at 30 degree to measure survival rate (CFU).



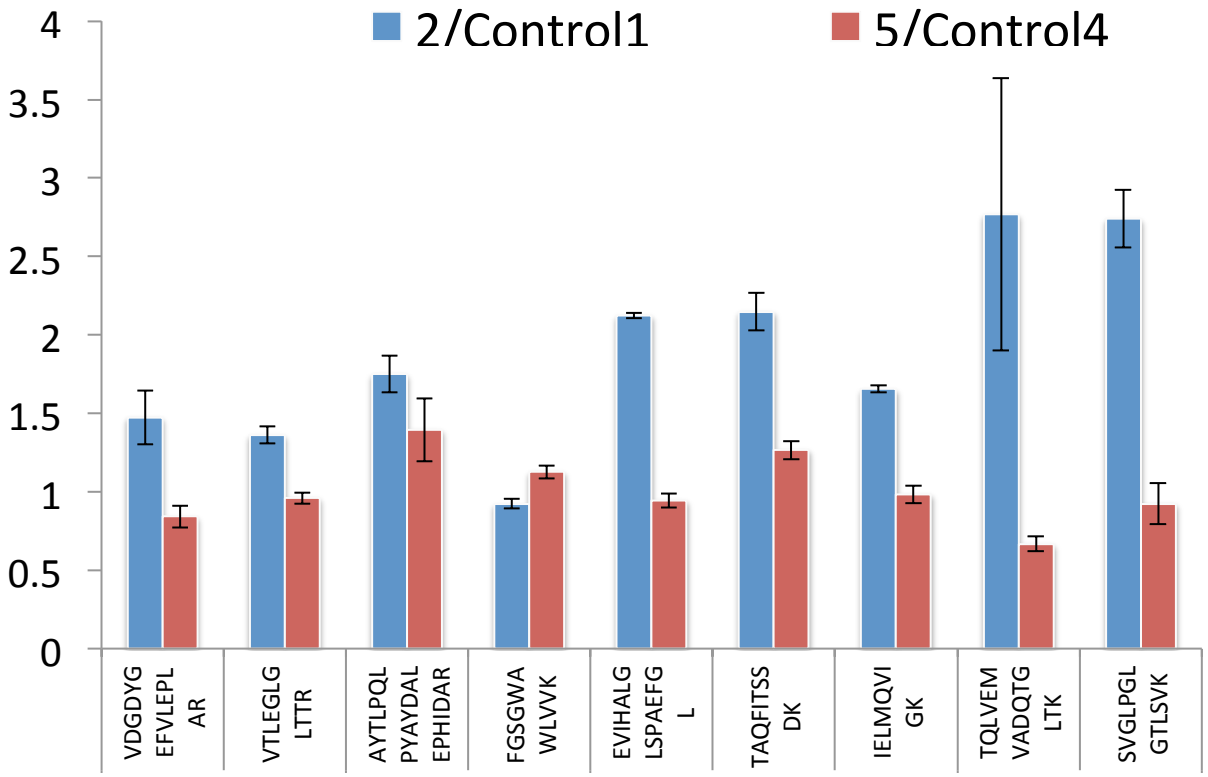
Total RNA and protein were prepared from recovered cells. Cells were sonicated and treated with lysozyme to obtain the protein lysate.

The protein lysates were digested with trypsin and analyzed with UPLC-MS/MS on the Orbitrap Elite.

Differential protein ID

High fold change proteins under 15 kGy irradiation in log phase	
Protein	Fold change
Serine esterase, GN=DR_0657	162
Succinate-semialdehyde dehydrogenase [NADP(+)], GN=ssdA	99
Fibronectin/fibrinogen-binding protein, GN=DR_0559	33
Alkaline shock protein-related protein, GN=DR_2068	14
N utilization substance protein B homolog, GN=nusB	14
Response regulator, GN=DR_0743	12
D-3-phosphoglycerate dehydrogenase, GN=DR_1291	10

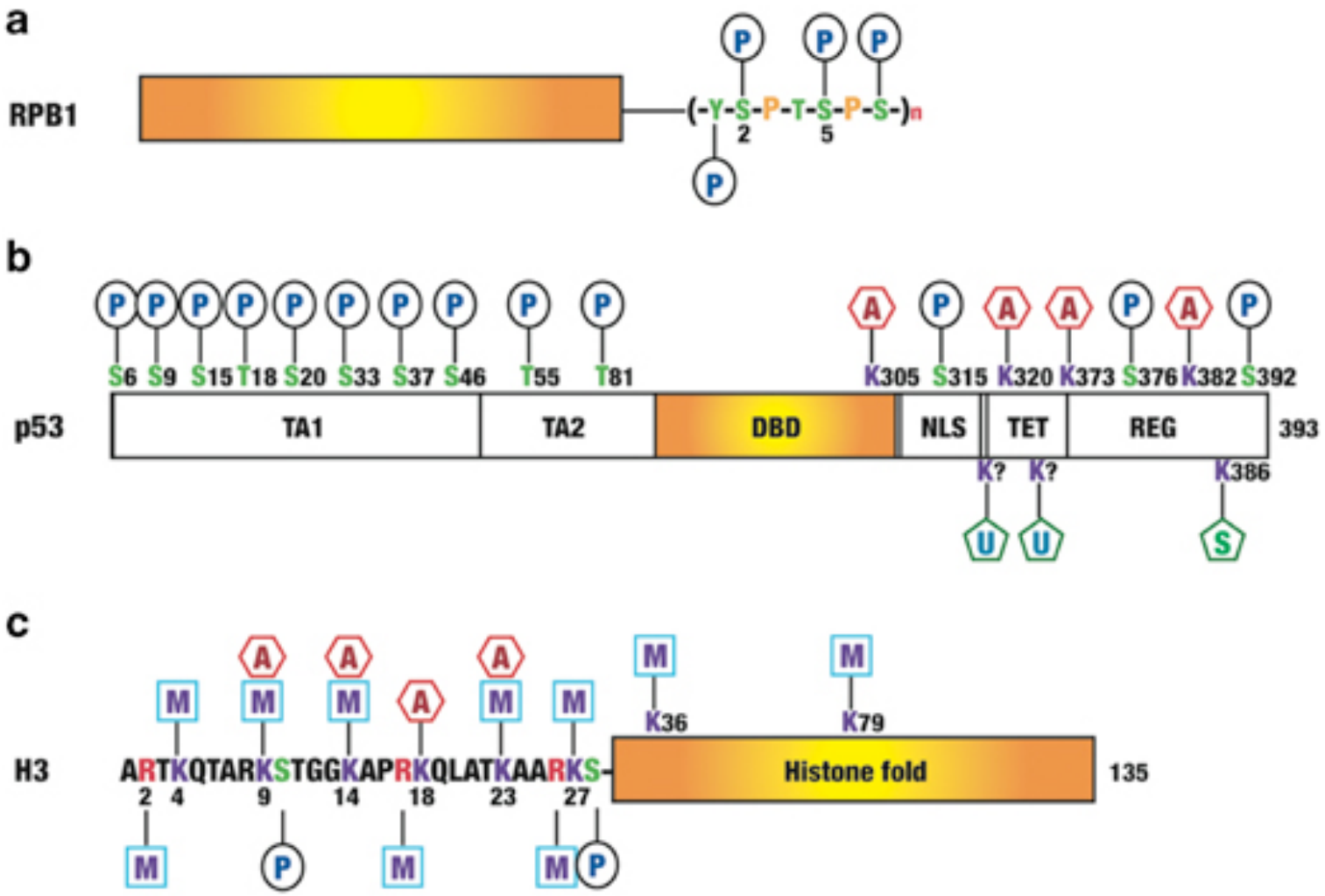
Quant with Synthetic Peptides



- *Samples were treated with low (2) and high (5) kGy*
- *Peak area from the targeted peptide is normalized against synthetic peptide*
- *Ratios obtained by comparing to non-irradiated controls*

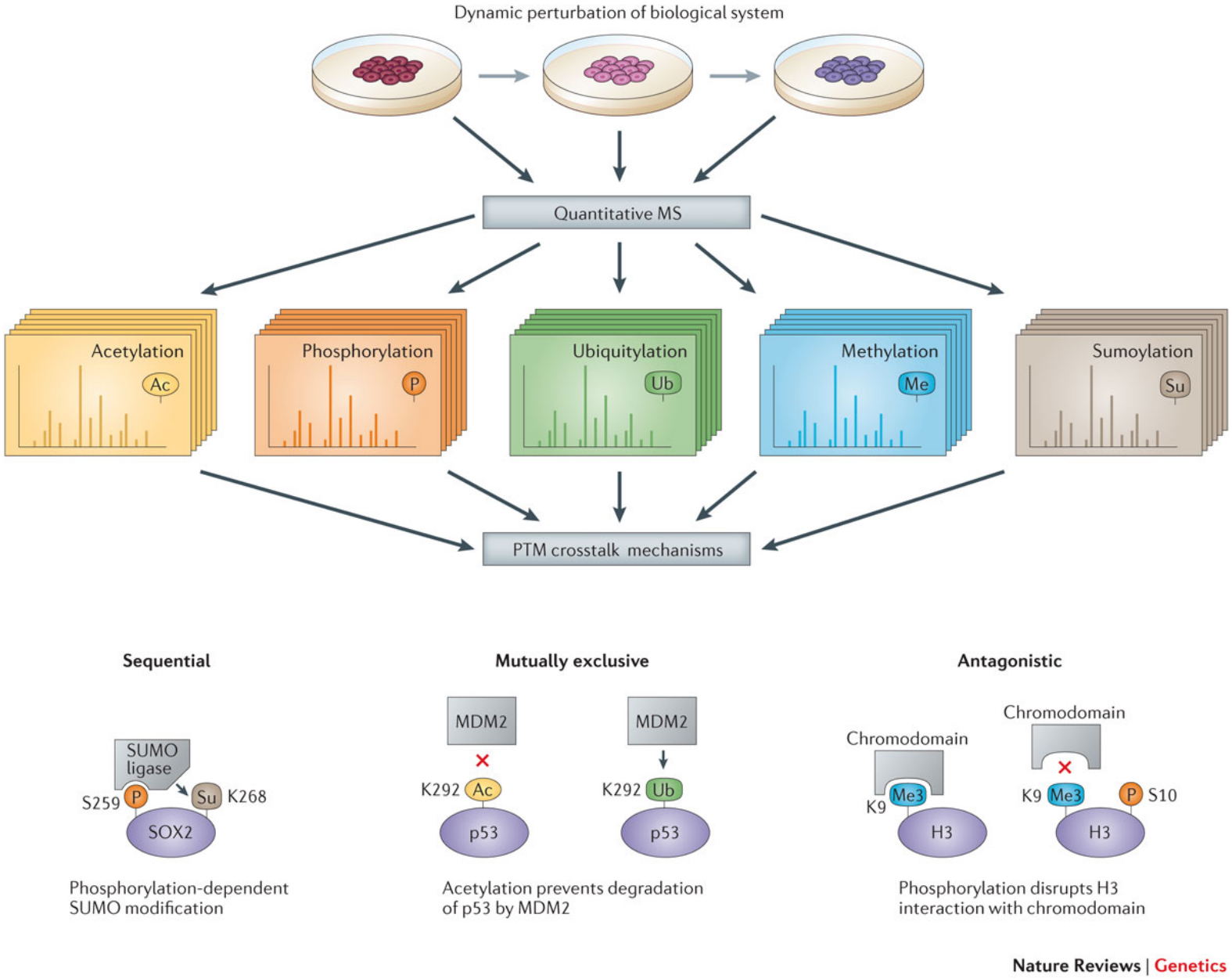
Post-Translational Modifications

Examples of Multiple PTMs per Protein



Modifications determine protein function, signaling, and localization

Xiang-Jiao Yang, **Multisite protein modification and intramolecular signaling** Oncogene (2005) 24, 1653–62

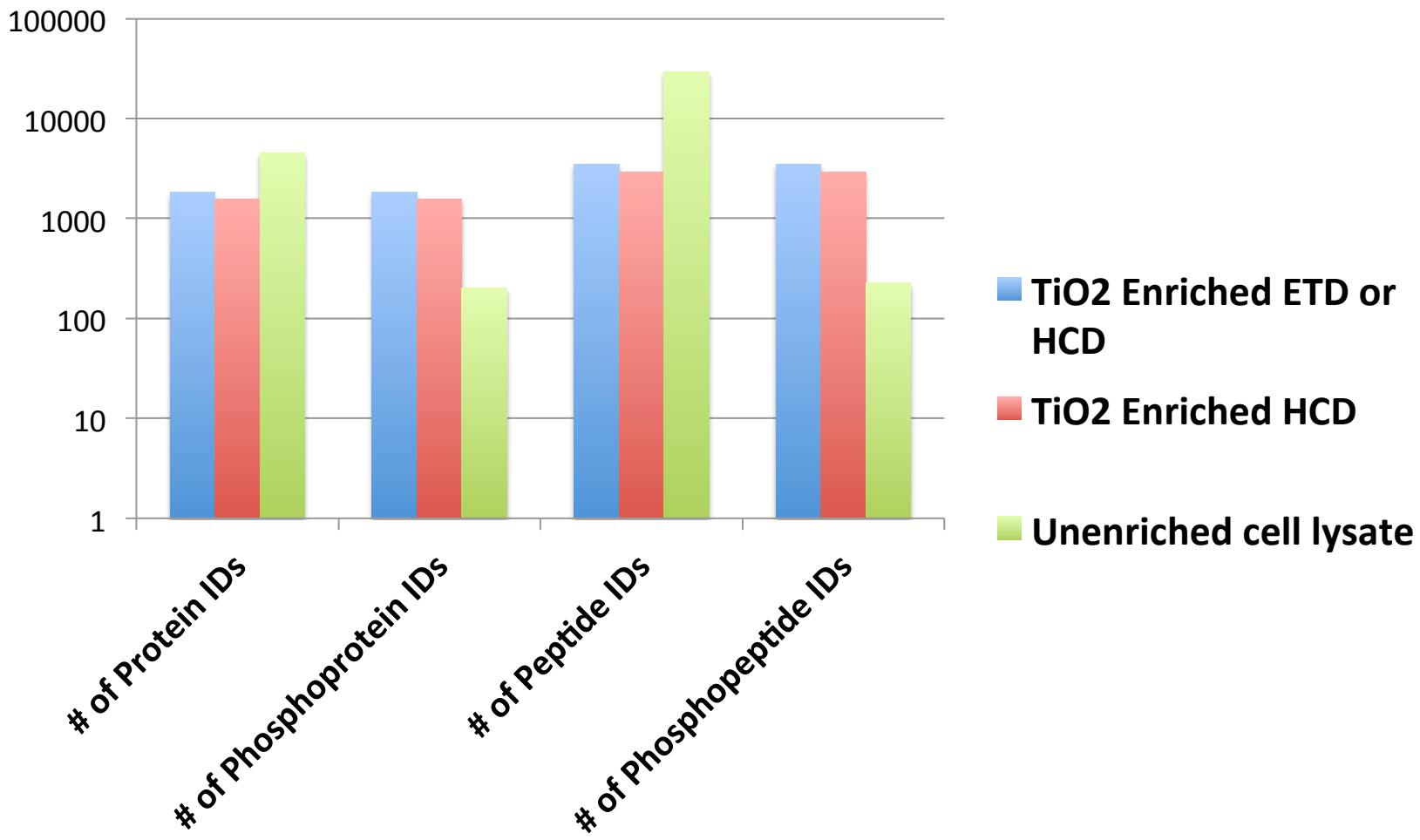


Altelaar et al. Nat Rev Genet. 2013 Jan;14(1):35-48

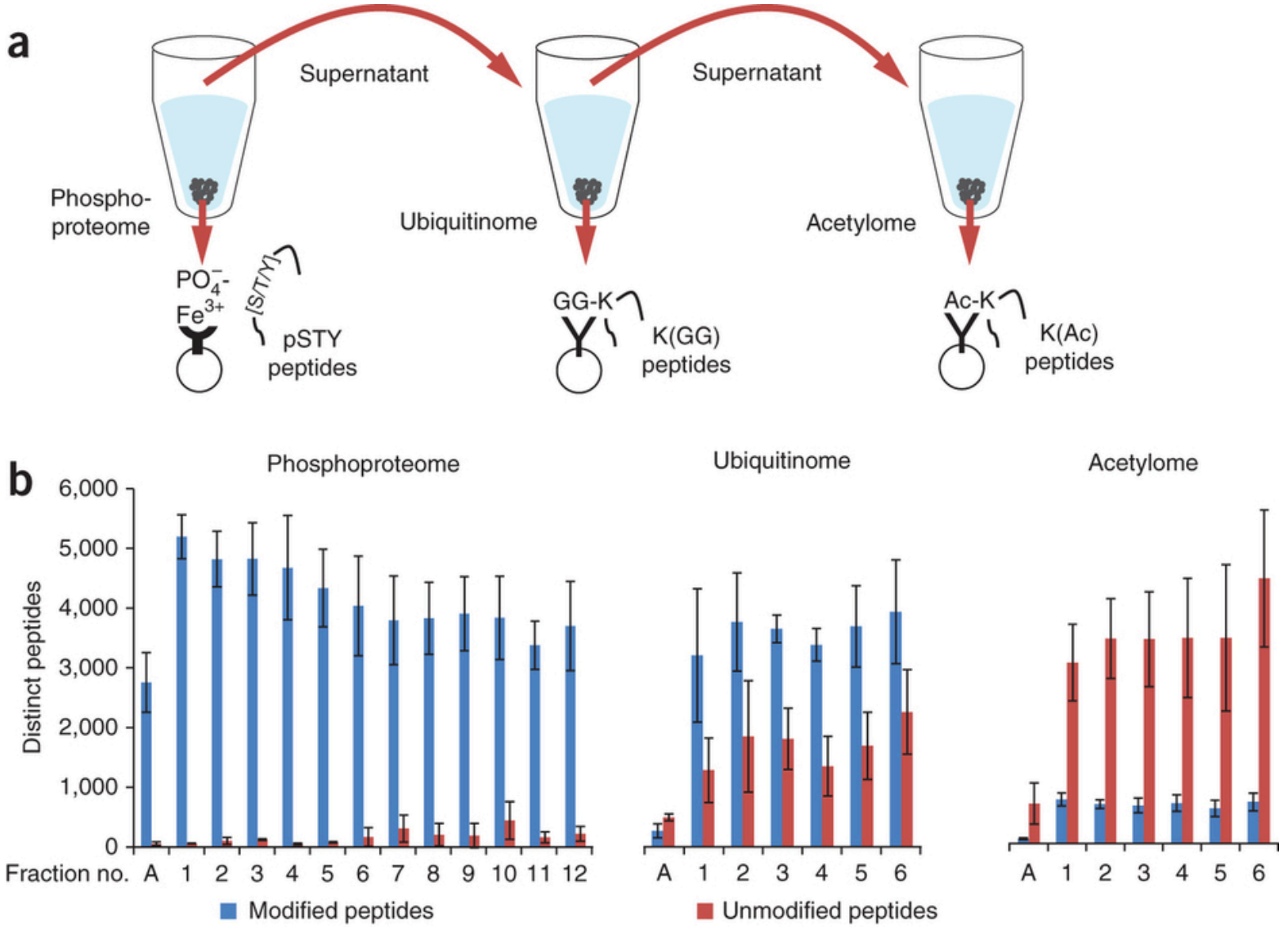
Detecting Modifications by MS

- Start with microgram levels of single protein or mg of lysate
- Use modification enrichment: affinity chromatography, antibody pulldown, biotinylation, click chemistry
- Purify protein/protein complex/organelle
- Use multiple proteases to increase coverage
- Try targeted MS/MS on modified peptide
- Use Ascore to asses site localization
- Validate with synthetic modified peptide standard or antibody

Phosphopeptide enrichment with TiO2 increases phosphopeptide identifications

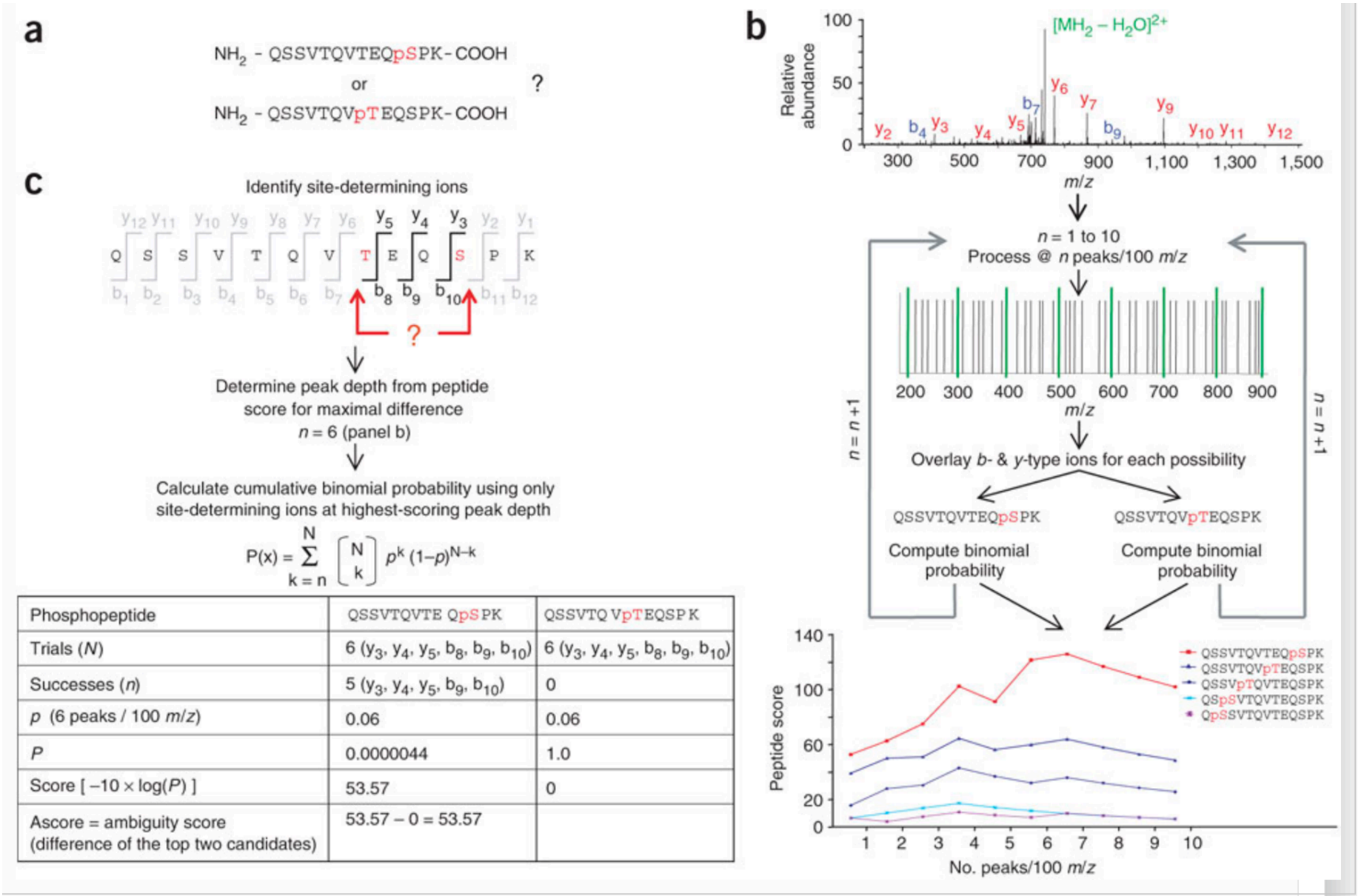


Serial Enrichment of PTMs with Basic RP Fractionation



Mertins P et al.. Nat Methods. 2013 10(7):634-7.

A Score for Localization of Modification



Phosphopeptide	QSSVTQVTE QpS PK	QSSVTQ VpTEQSP K
Trials (*N*)	6 (y_3, y_4, y_5, b_8, b_9, b_{10})	6 (y_3, y_4, y_5, b_8, b_9, b_{10})
Successes (*n*)	5 (y_3, y_4, y_5, b_9, b_{10})	0
p (6 peaks / 100 *m/z*)	0.06	0.06
P	0.0000044	1.0
Score [$-10 \times \log(P)$]	53.57	0
Ascore = ambiguity score (difference of the top two candidates)	53.57 – 0 = 53.57	

Beausoleil SA, Nat Biotechnol. 2006 Oct;24(10):1285-92

Scaffold PTM

- Open Scaffold PTM program
- If asked about database access, cancel
- Run Demo Tutorial 1 Single MS Sample
- Select PTM List on left sidebar, scroll through results
- Select BC11B Go to Proteins View
- Lower right pane is Spectrum+A score
- Go to Motif View

Slide Acknowledgements

http://www.matrixscience.com/help_index.html

http://proteome-software.wikispaces.com/Proteomics
Brian Searle: Interpreting MS/MS Proteomics Results

Thermo and Piercenet websites

Joseph A. Loo, UCLA, ppt entitled "*Mass Spectrometry for Protein Quantification and Identification of Posttranslational Modifications*"

Olga Vitek, US HUPO 2016, Statistics for (Targeted, Label-Free) Proteomics

References

- UT Austin Proteomics Facility Proteomics Educational Links https://wikis.utexas.edu/display/proteomicscore/Proteomics+educational+links
  Links to webpages, lectures and videos on mass spectrometry, protein identification by database search, and proteomics applications

- Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016 Sep 15;537(7620):347-55.
- Bantscheff et al., “Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present” Anal Bioanal Chem (2012) 404:939-965
- Egertson et al., Multiplexed MS/MS for improved data-independent acquisition Nature Methods 10, 744–746 (2013) doi:10.1038/nmeth.2528
- Doerr A. DIA mass spectrometry. Nature Methods 12, 35 (2015) doi: 10.1038/nmeth.3234