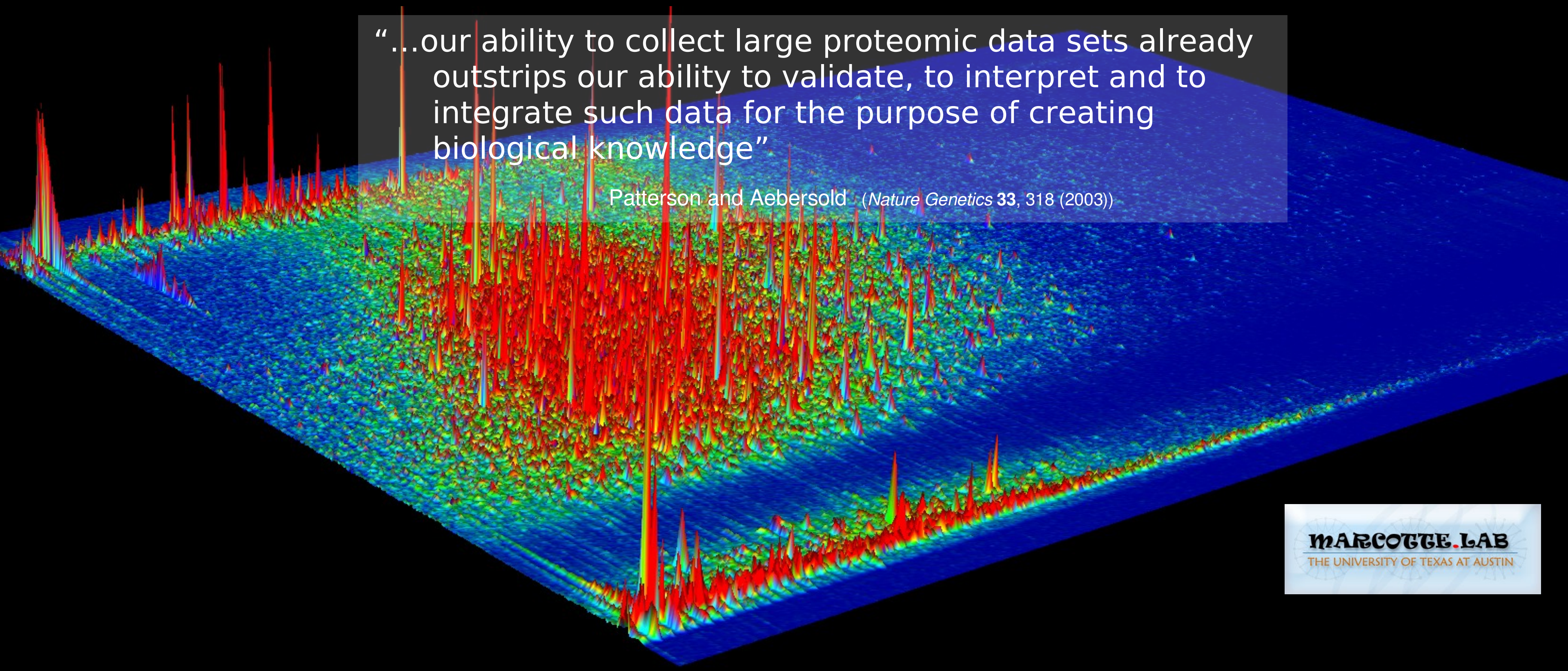
Mass Spectrometry Proteomics for the Computational Biologist

December 1, 2006

John T. Prince



> "...our ability to collect large proteomic data sets already outstrips our ability to validate, to interpret and to integrate such data for the purpose of creating biological knowledge"
>
> Patterson and Aebersold (*Nature Genetics* **33**, 318 (2003))

MARCOTTE.LAB
THE UNIVERSITY OF TEXAS AT AUSTIN

Mass Spectrometry (MS) Proteomics

Needs Computational Biologists

"…our ability to collect large proteomic data sets already outstrips our ability to validate, to interpret and to integrate such data for the purpose of creating biological knowledge"

- Patterson and Aebersold (*Nature Genetics* **33**, 318 (2003))

MS Proteomics

- How?
- Data?
- Problems?

Why Proteomics?

and not just Transcriptomics

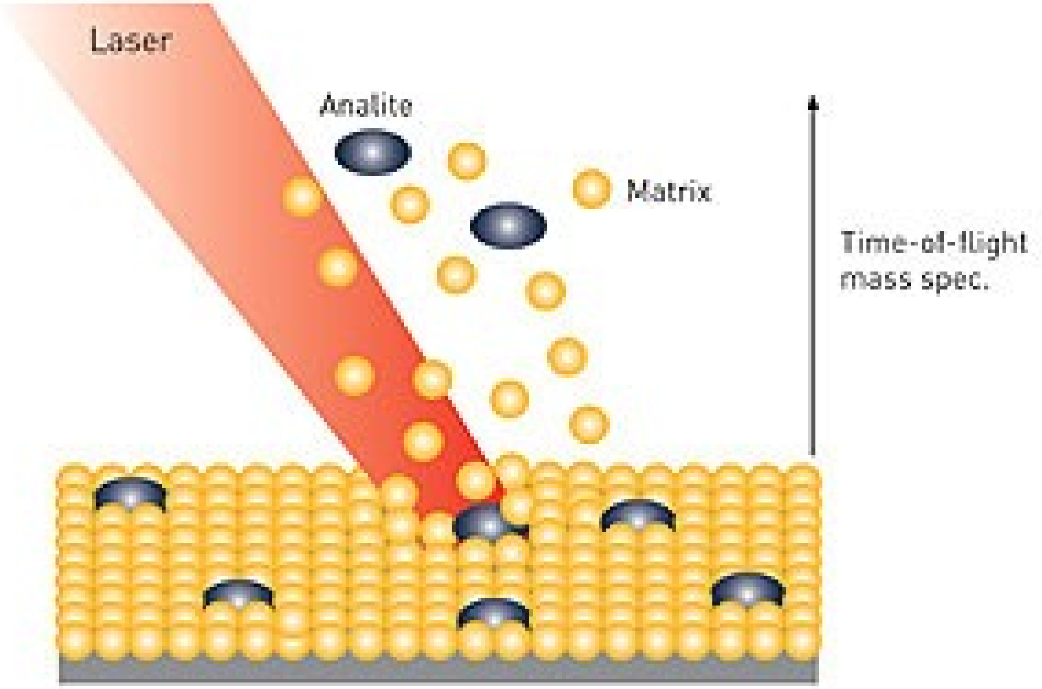
- Proteins are the actual players
- mRNA not necessarily proportional to protein level
  - translational control
  - degradation
- Post-translational modifications alter cell state
- Cellular localization

Mass Spec (Proteomics)

- Ionization
  - MALDI
  - ESI
- m/z Analysis
  - TOF
  - Quadrupole
  - Ion Trap
  - FTICR
  - Orbitrap

MALDI

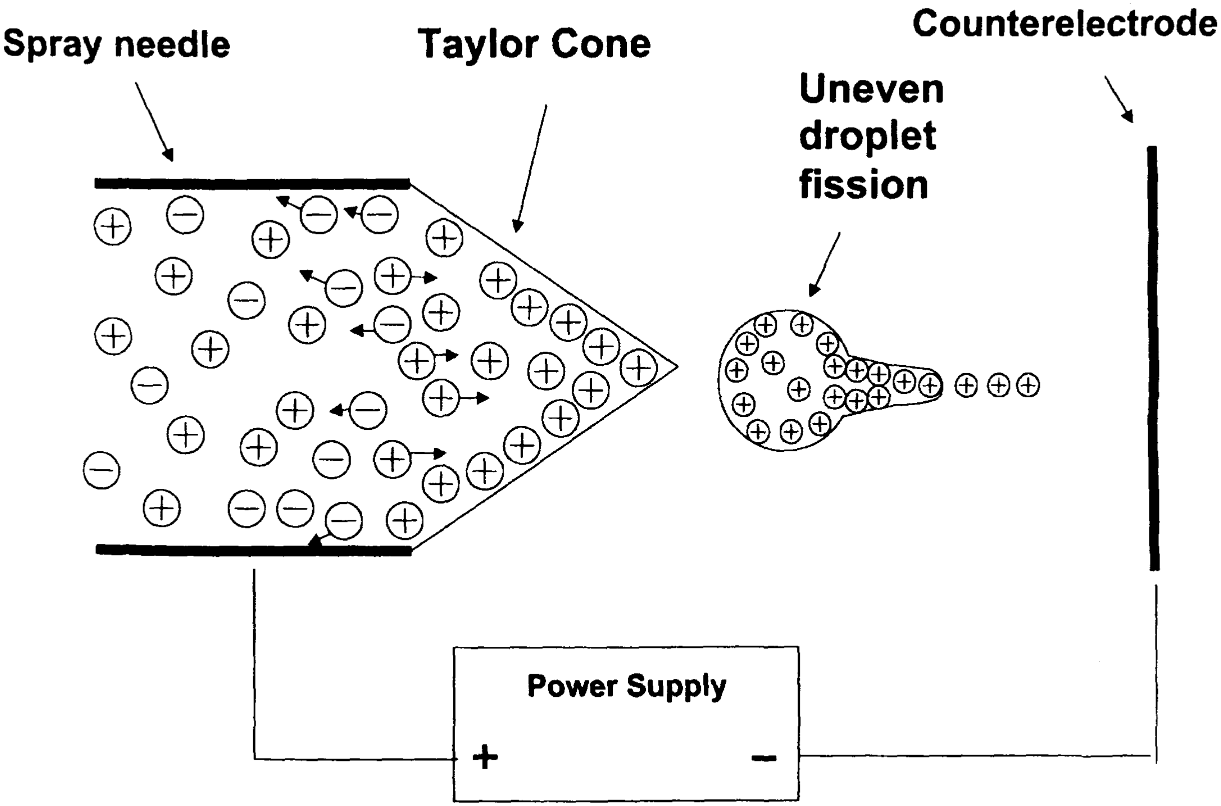
Matrix Assisted Laser Desorption Ionization



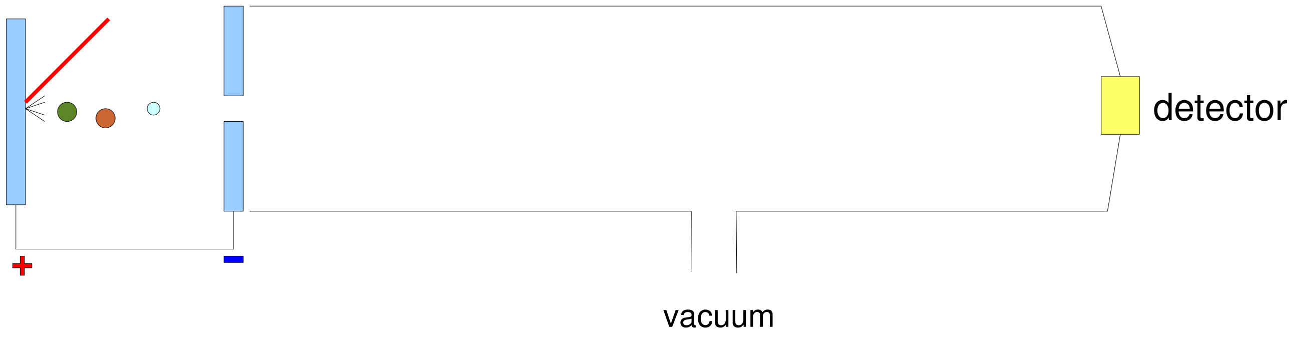
http://www.eurogentec.com/module/images2/p23_3.jpg

ESI
Electrospray Ionization

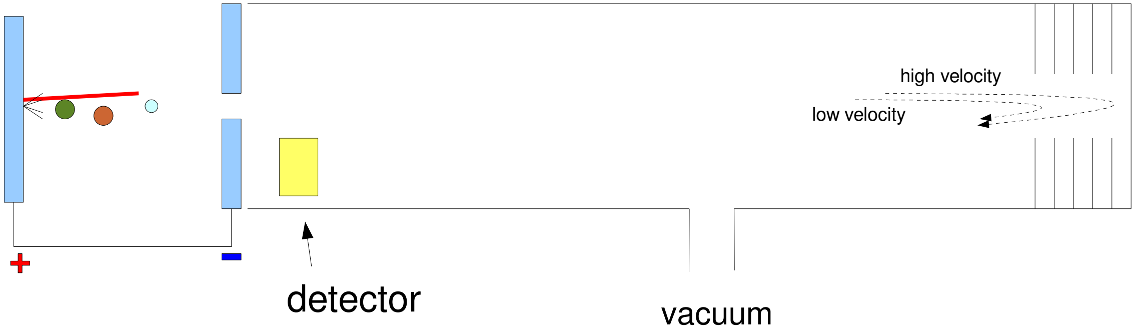
http://www.phoenix-st.com/images/splash2.jpg



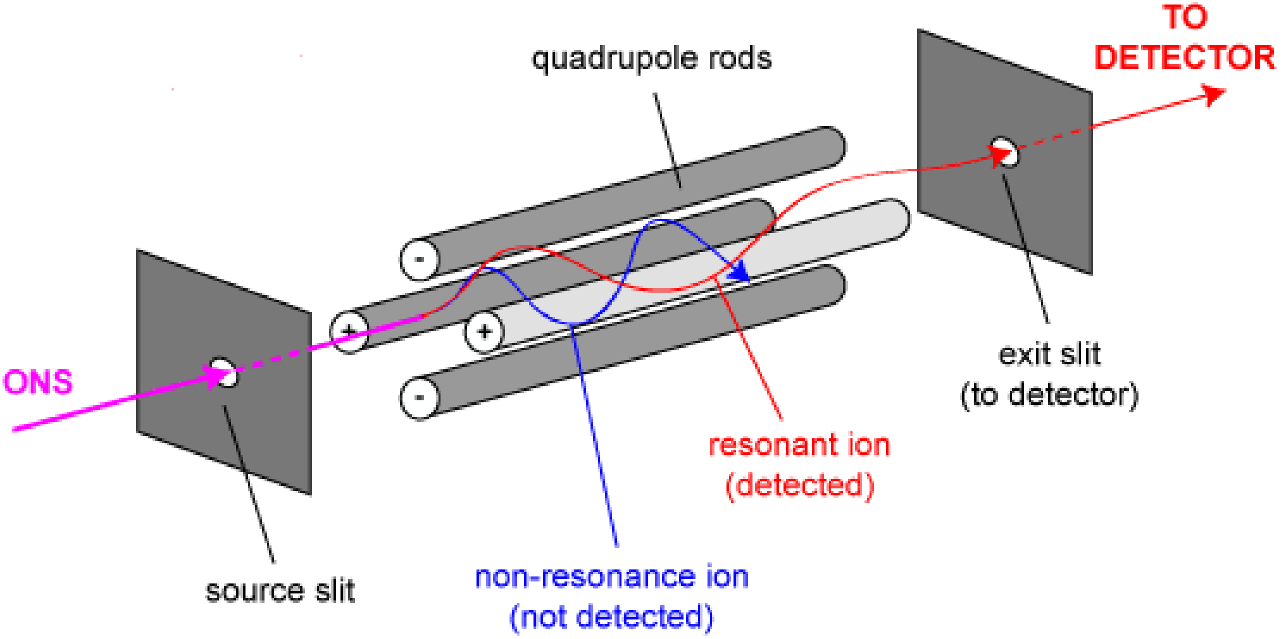
TOF
Time of Flight



TOF (reflectron)

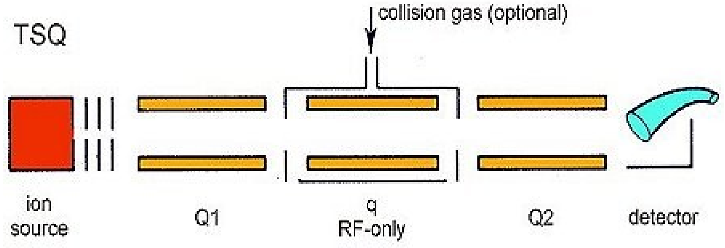


Q
(e.g., Q-TOF, QQQ)Quadrupole



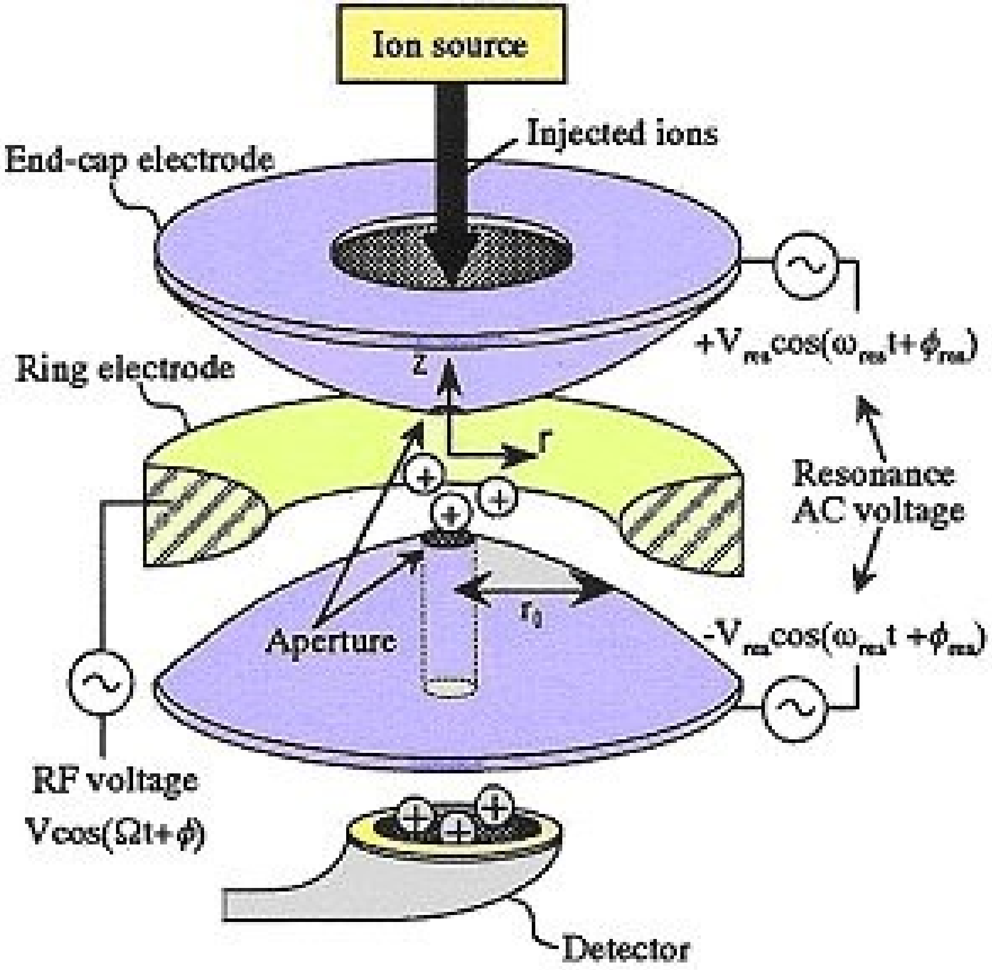
http://www.bris.ac.uk/nerclsmsf/images/quadrupole.gif

QQQ



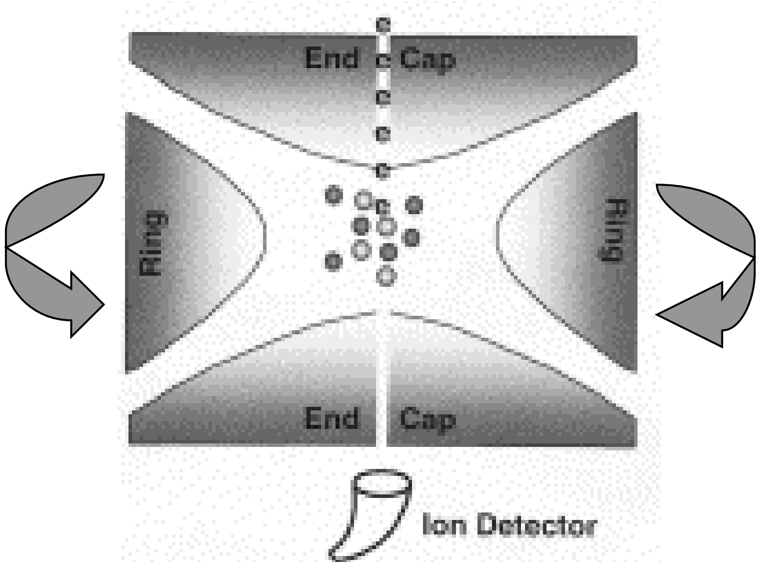
http://www.rzuser.uni-heidelberg.de/~bl5/ency/pics/t_tsq1.jpg

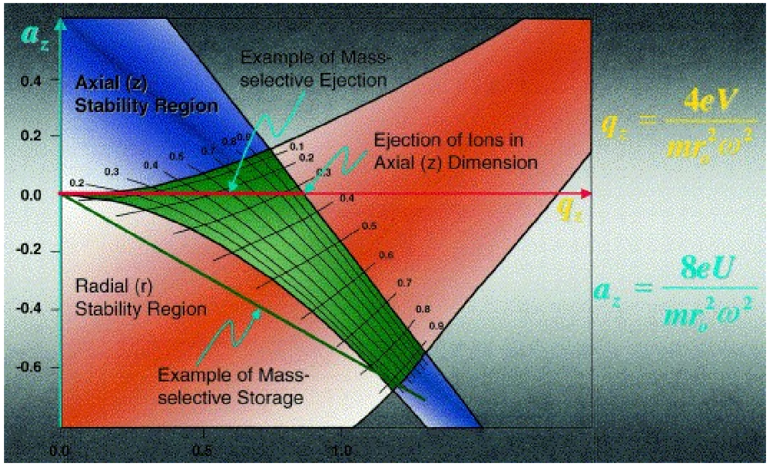
Quadrupole Ion Trap



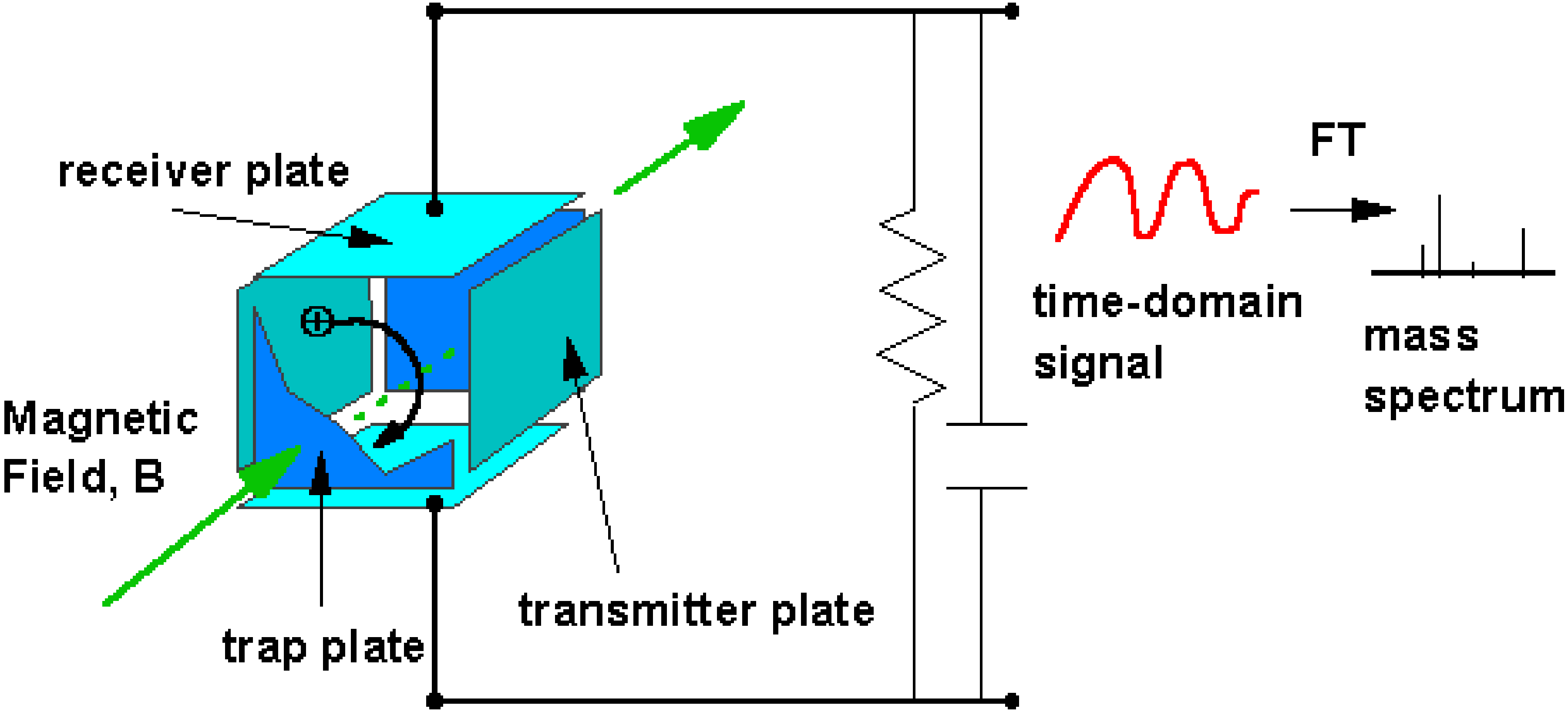
http://www.rzuser.uni-heidelberg.de/~bl5/ency/pics/q_trap01.jpg

K. Yoshinari, Rapid Commun. Mass Spectrom. 14, 215-223 (2000)





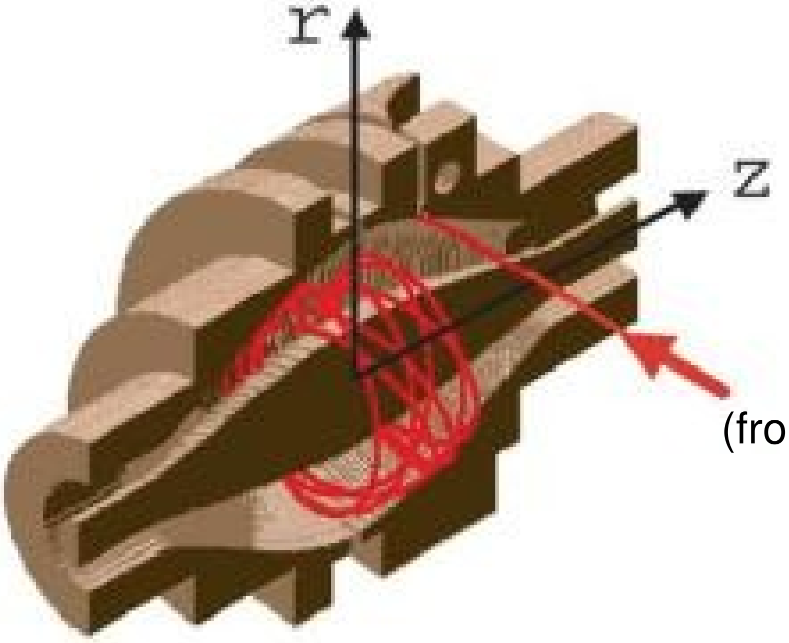
FT-ICR
Fourier Transform Ion Cyclotron Resonance



http://www.ivv.fraunhofer.de/ms/ms-analyzers.html

FT-Orbi

Orbitrap



(from Linear Ion Trap via C trap)

http://www.spectroscopynow.com/ftp_images/orbitrap_0505.jpg

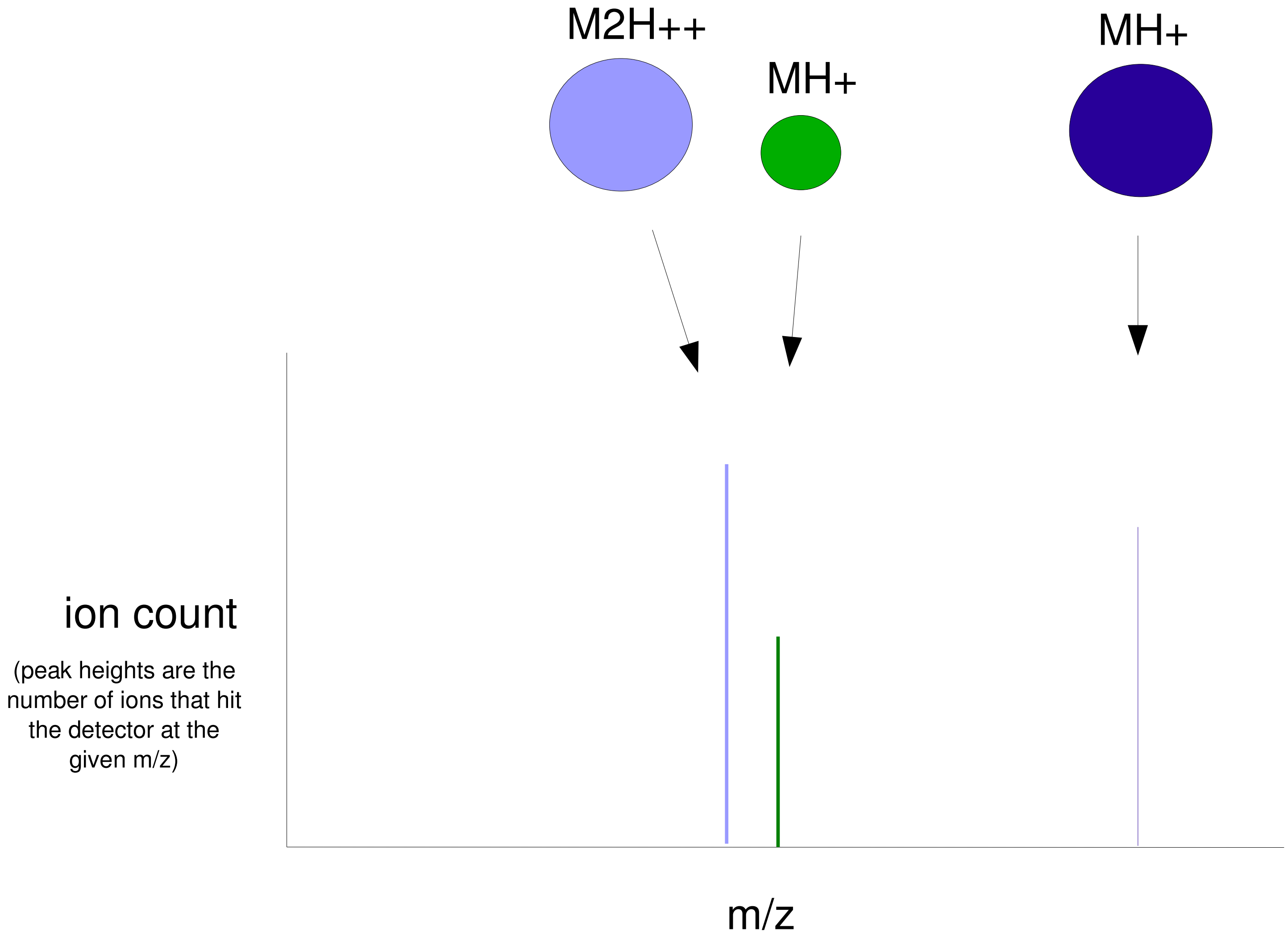
Mass Spectrometry (Proteomics)

- Ionization
  - ESI (Electrospray Ionization)
  - MALDI (Matrix Assisted Laser Desorption Ionization)
- m/z Analysis
  - TOF (Time of Flight)
  - Q ([e.g. Q-TOF] Quadrupole)
  - Ion Trap
  - FTICR (Fourier Transform Ion Cyclotron Resonance)
  - Orbitrap

Data

- Spectrum
- ESI Protein Spectrum
- 2D MALDI imaging
- PMF (peptide mass fingerprinting)
- LC-MS
- Peptide Fragmentation
- MudPIT

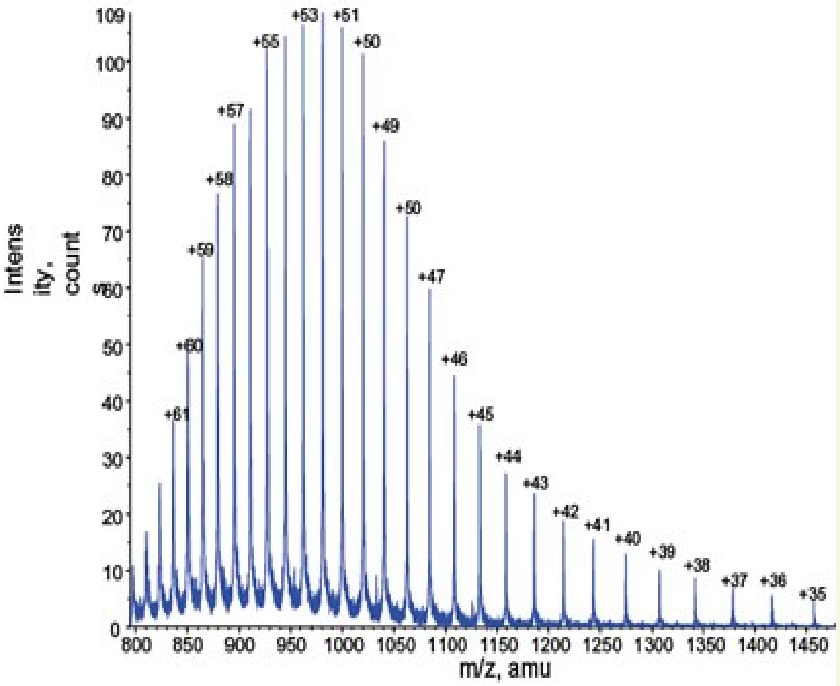
Spectrum



M2H++
MH+
MH+
ion count
(peak heights are the number of ions that hit the detector at the given m/z)
m/z

Why Not Proteins?

Multiple Charge States (ESI)



http://www-methods.ch.cam.ac.uk/siteimages/sw3.jpg

PTMs (Post-Translational Modifications)

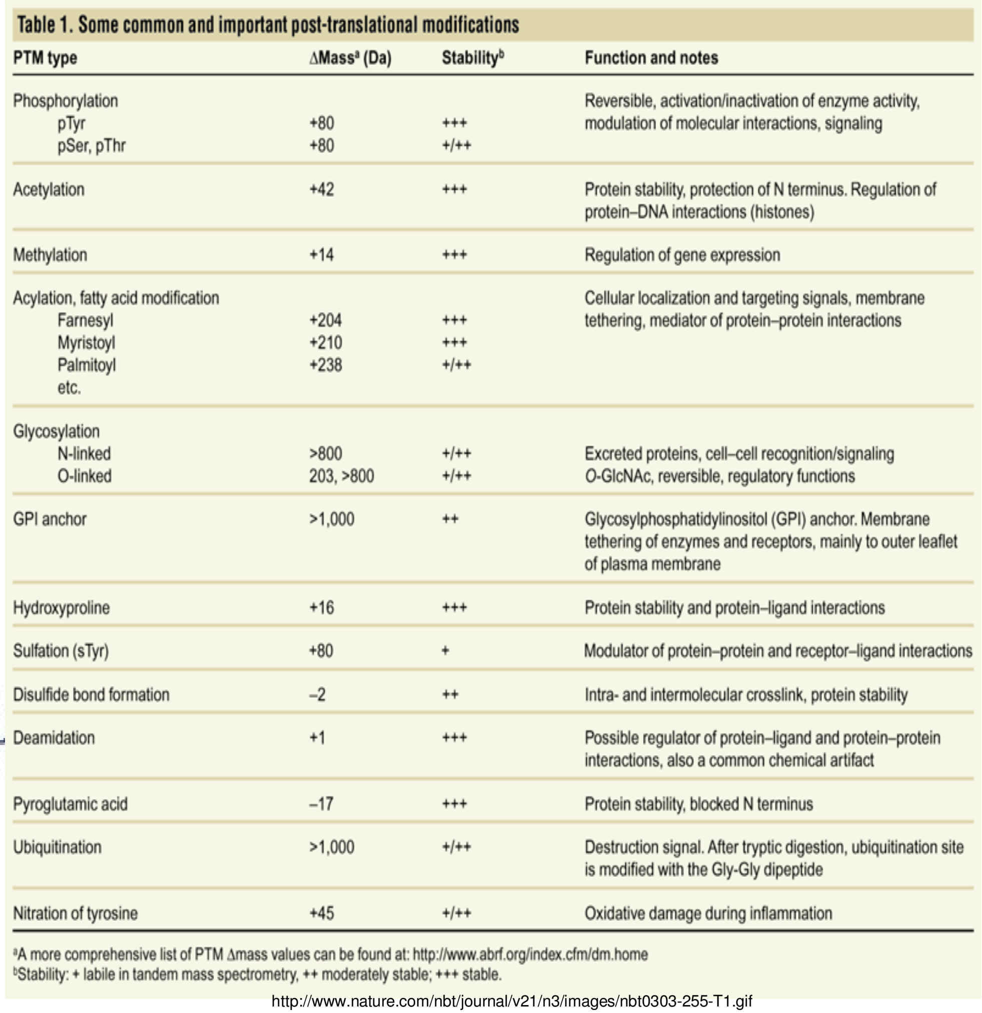


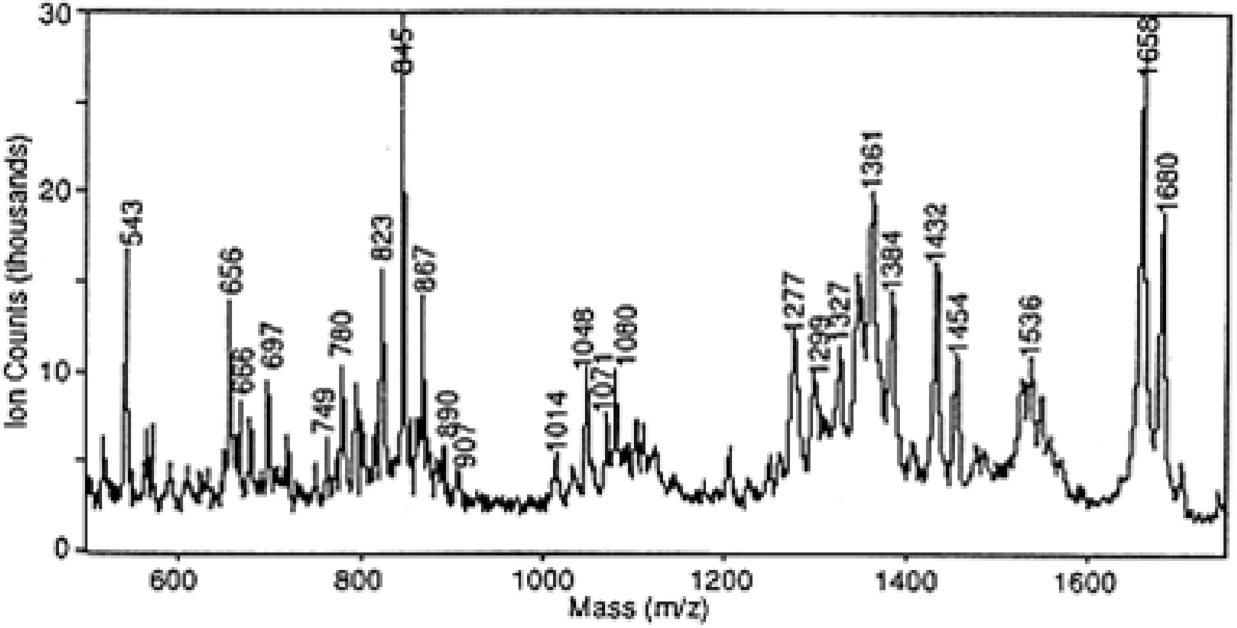
Table 1. Some common and important post-translational modifications

PTM type	ΔMass[a] (Da)	Stability[b]	Function and notes
Phosphorylation			Reversible, activation/inactivation of enzyme activity, modulation of molecular interactions, signaling
pTyr	+80	+++	
pSer, pThr	+80	+/++	
Acetylation	+42	+++	Protein stability, protection of N terminus. Regulation of protein–DNA interactions (histones)
Methylation	+14	+++	Regulation of gene expression
Acylation, fatty acid modification			Cellular localization and targeting signals, membrane tethering, mediator of protein–protein interactions
Farnesyl	+204	+++	
Myristoyl	+210	+++	
Palmitoyl	+238	+/++	
etc.			
Glycosylation			
N-linked	>800	+/++	Excreted proteins, cell–cell recognition/signaling
O-linked	203, >800	+/++	*O*-GlcNAc, reversible, regulatory functions
GPI anchor	>1,000	++	Glycosylphosphatidylinositol (GPI) anchor. Membrane tethering of enzymes and receptors, mainly to outer leaflet of plasma membrane
Hydroxyproline	+16	+++	Protein stability and protein–ligand interactions
Sulfation (sTyr)	+80	+	Modulator of protein–protein and receptor–ligand interactions
Disulfide bond formation	–2	++	Intra- and intermolecular crosslink, protein stability
Deamidation	+1	+++	Possible regulator of protein–ligand and protein–protein interactions, also a common chemical artifact
Pyroglutamic acid	–17	+++	Protein stability, blocked N terminus
Ubiquitination	>1,000	+/++	Destruction signal. After tryptic digestion, ubiquitination site is modified with the Gly-Gly dipeptide
Nitration of tyrosine	+45	+/++	Oxidative damage during inflammation

[a] A more comprehensive list of PTM Δmass values can be found at: http://www.abrf.org/index.cfm/dm.home
[b] Stability: + labile in tandem mass spectrometry, ++ moderately stable; +++ stable.

http://www.nature.com/nbt/journal/v21/n3/images/nbt0303-255-T1.gif

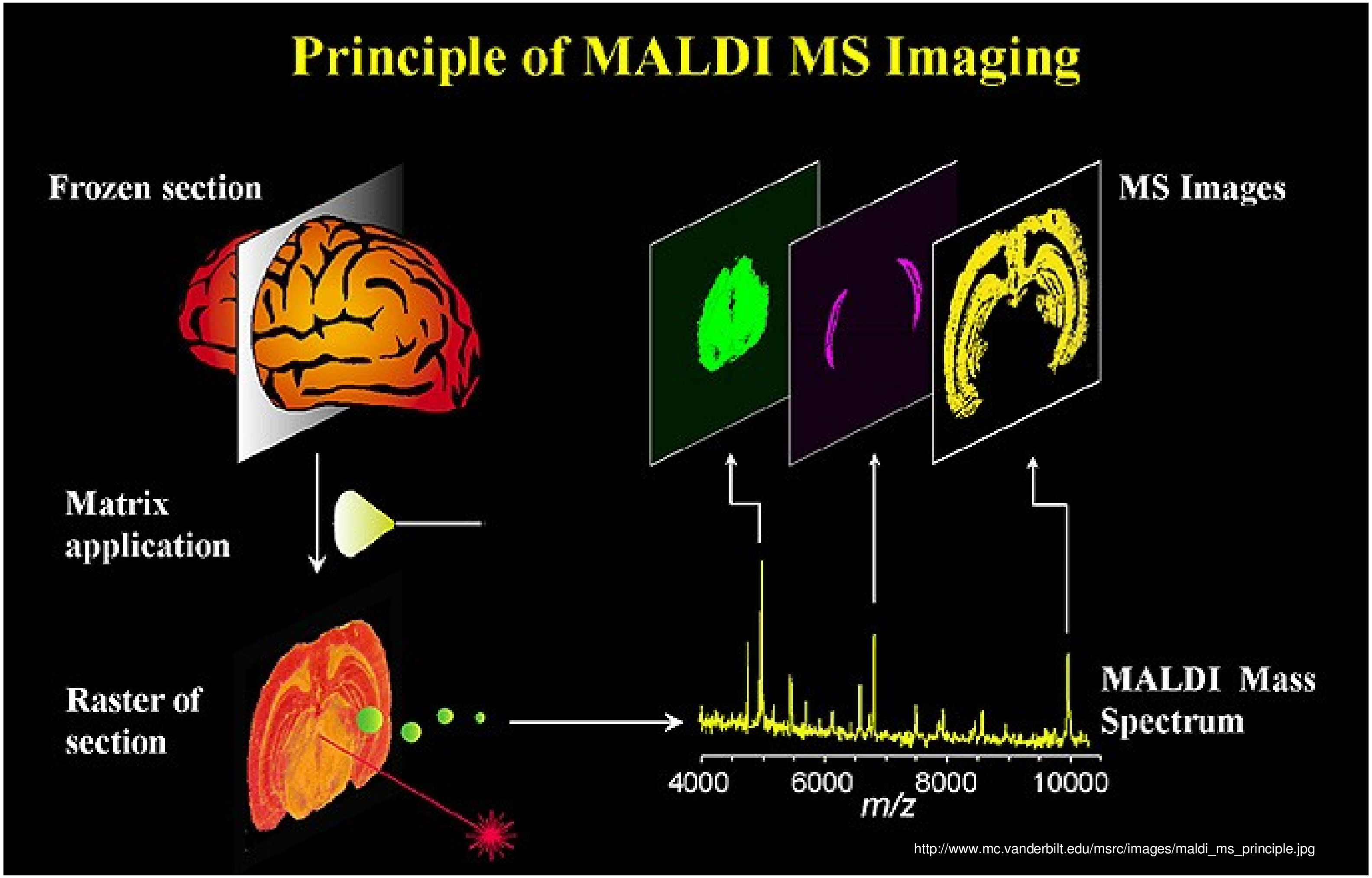
MALDI on Biological Sample



- Signal Processing
- Classification Analysis

http://www.mcb.mcgill.ca/~hallett/GEP/PLecture4/image002.gif

MALDI-TOF Application

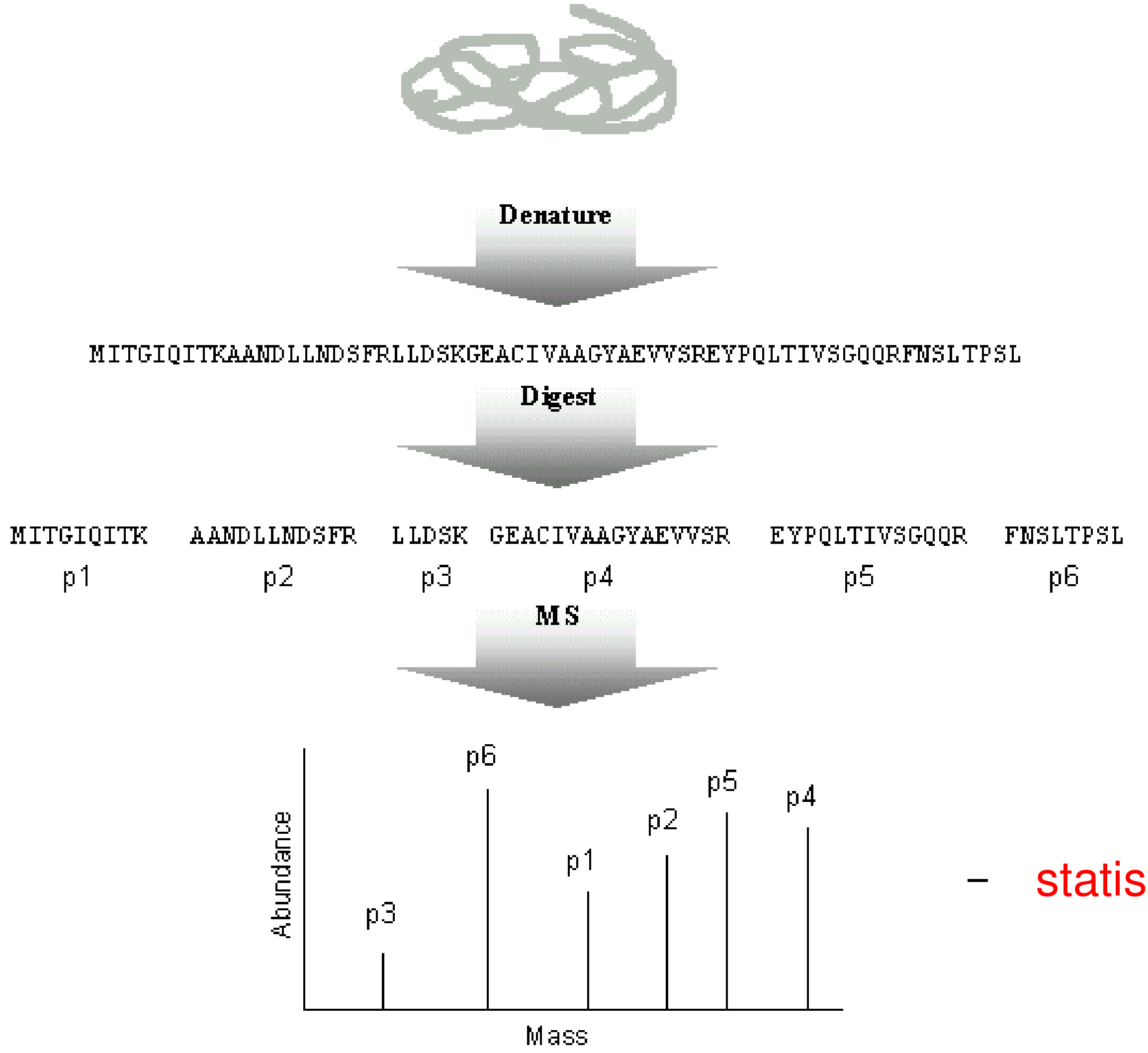


http://www.mc.vanderbilt.edu/msrc/images/maldi_ms_principle.jpg

- organize data
- integrate data
- mine data

PMF

Peptide Mass Fingerprinting

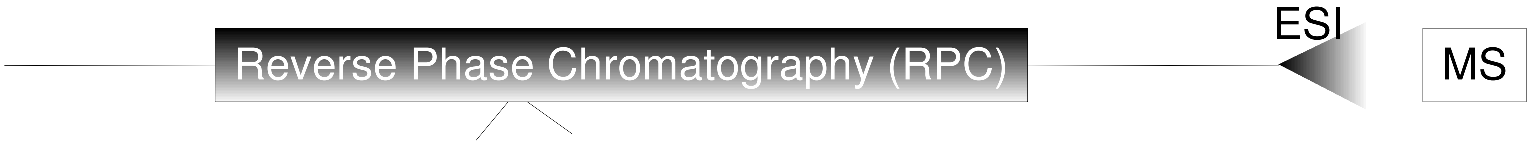


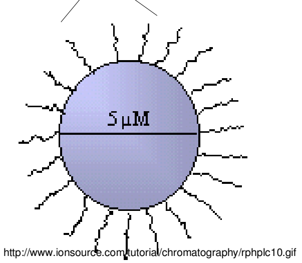
– statistical validation

http://www.spectroscopynow.com/ftp_images/2Fig3.gif

LC-MS

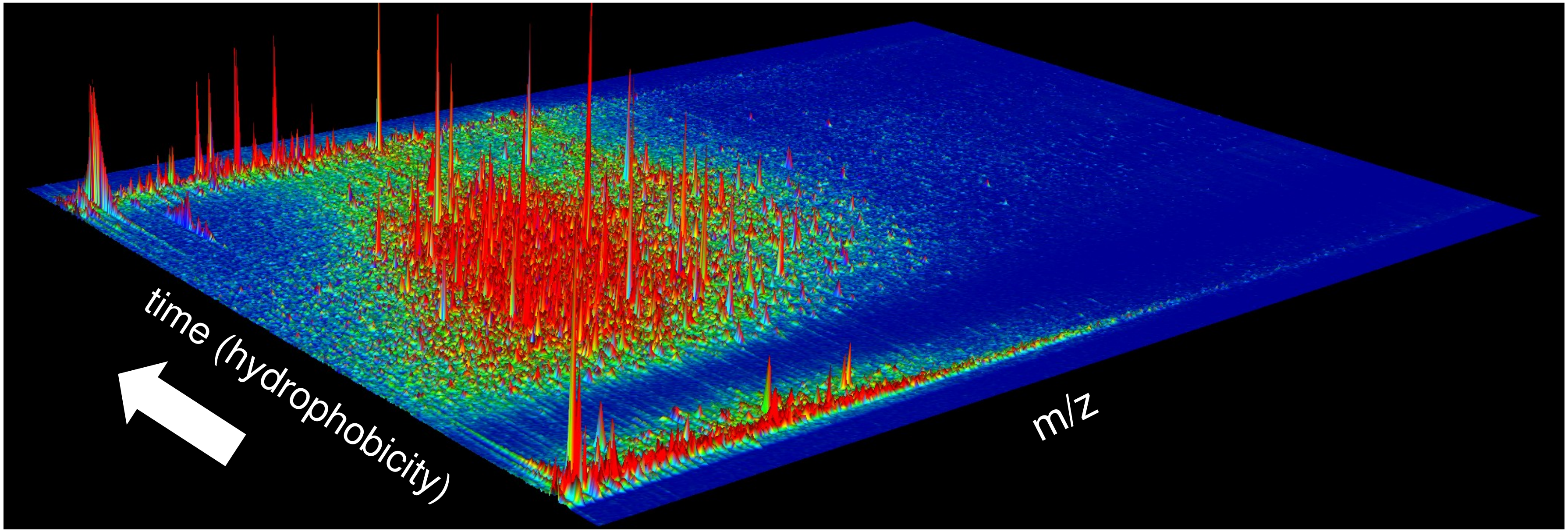
Liquid Chromatography MS



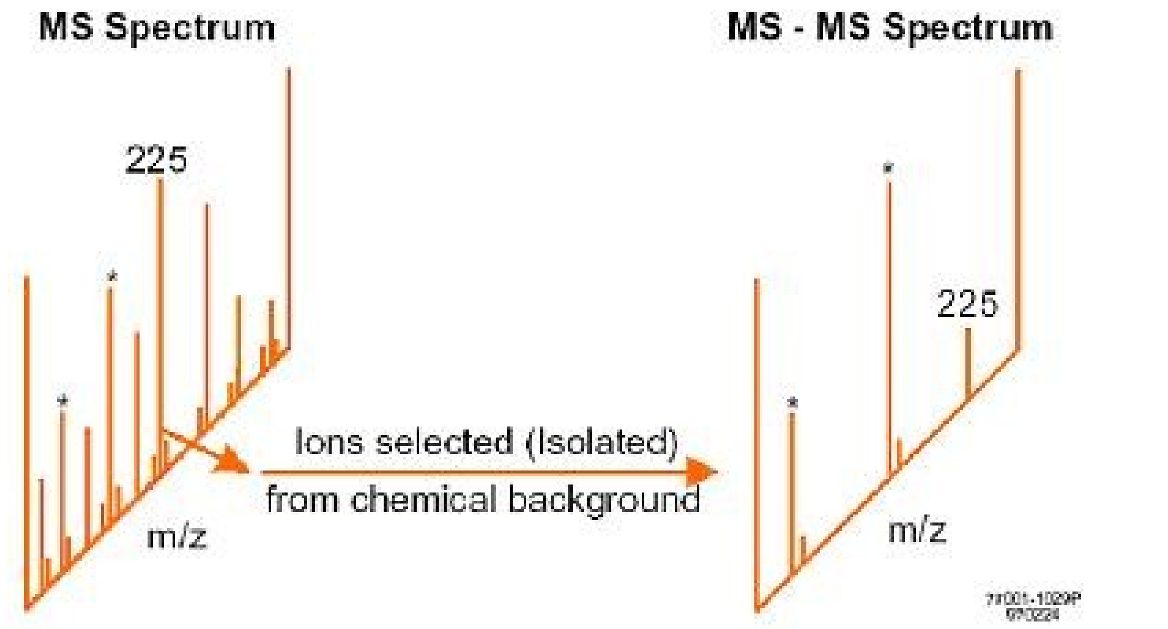


http://www.ionsource.com/tutorial/chromatography/rphplc10.gif

- elution prediction
- registration



MS/MS (Peptide Fragmentation)

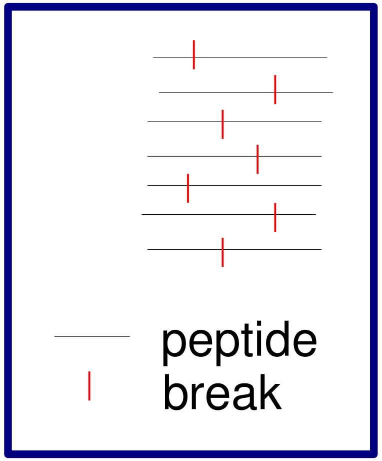


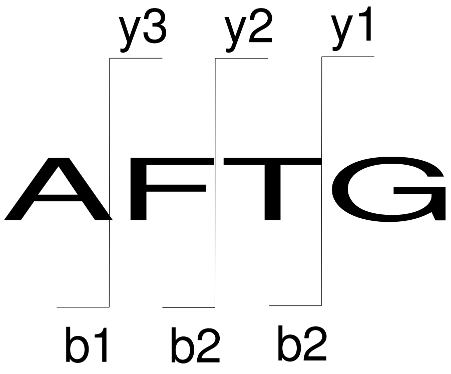
1st stage of MS-MS

2nd stage of MS-MS
(Collision-induced dissociation)

http://www.mcb.mcgill.ca/~hallett/GEP/PLecture5/PLecture5.html

Peptide Fragmentation (MS/MS)





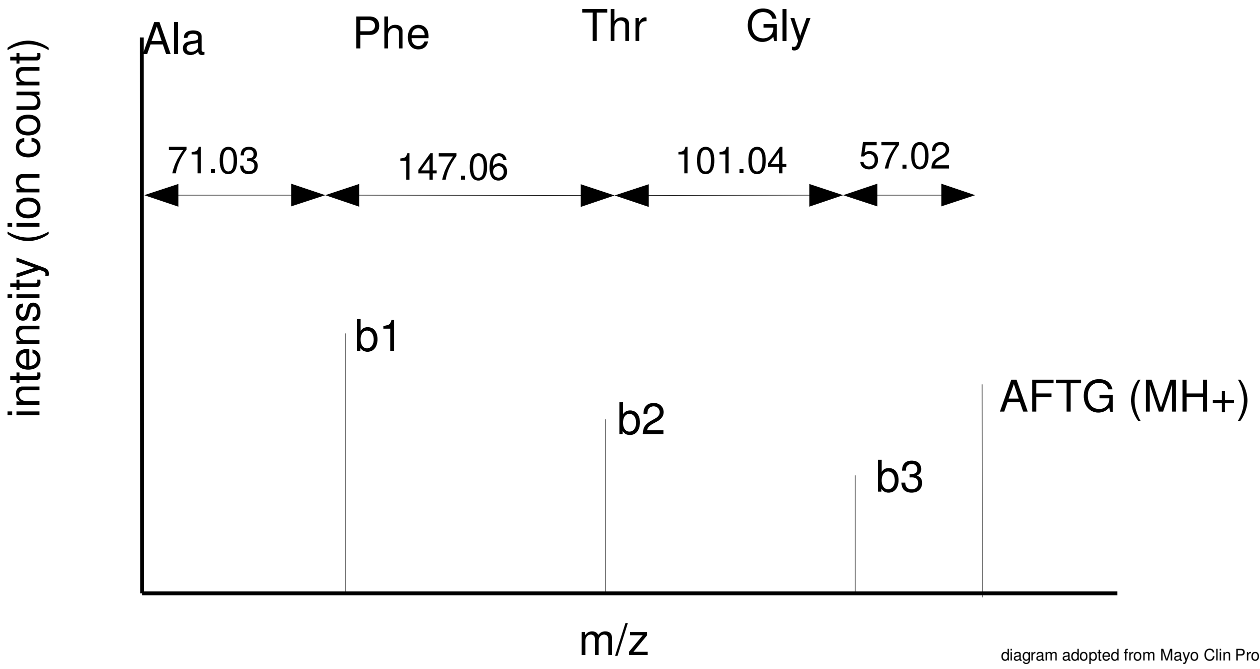
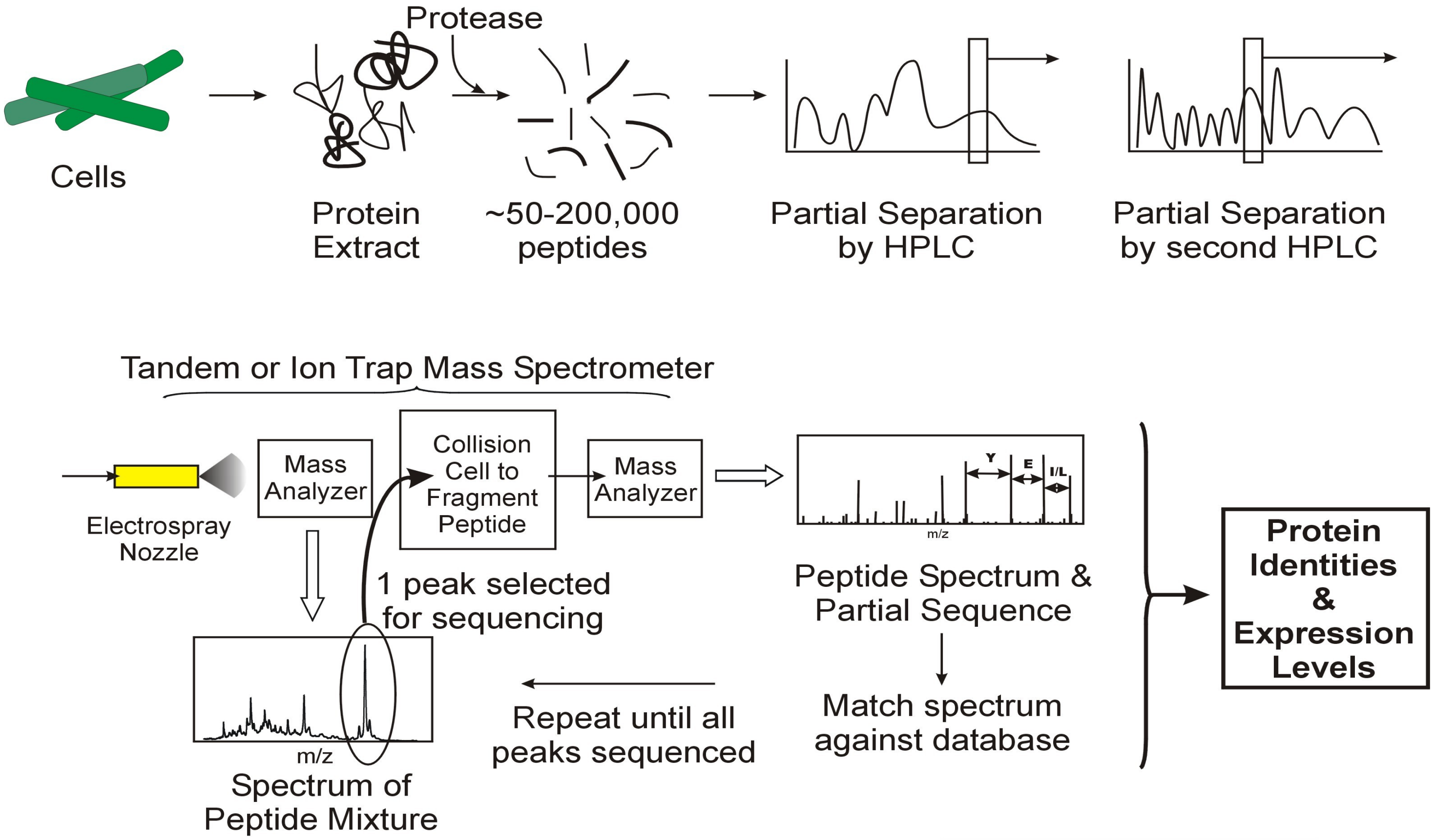


diagram adopted from Mayo Clin Proc. 2002;77:1185-1196

Shotgun Proteomics



Cells

Protease

Protein Extract

~50-200,000 peptides

Partial Separation by HPLC

Partial Separation by second HPLC

Tandem or Ion Trap Mass Spectrometer

Electrospray Nozzle

Mass Analyzer

Collision Cell to Fragment Peptide

Mass Analyzer

1 peak selected for sequencing

Spectrum of Peptide Mixture

Repeat until all peaks sequenced

Peptide Spectrum & Partial Sequence

Match spectrum against database

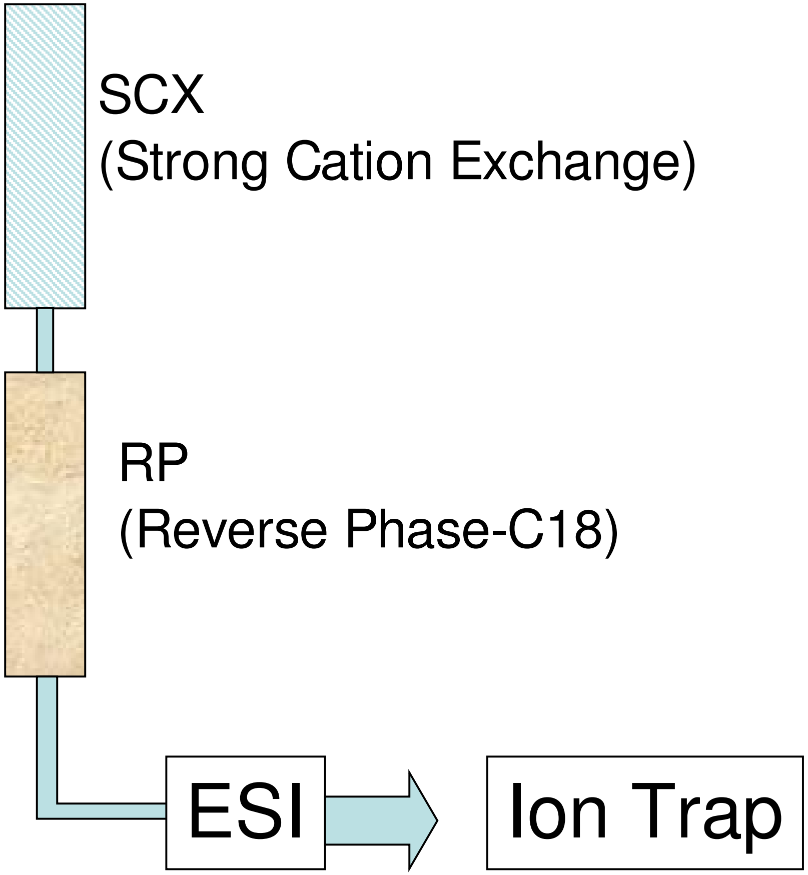
Protein Identities & Expression Levels

> 3 million data points per experiment

quantitation
peptide fragmentation prediction
spectra comparison metrics
peptides to proteins
integrating bayesian priors

MuDPIT

Multidimensional Protein Identification Technology

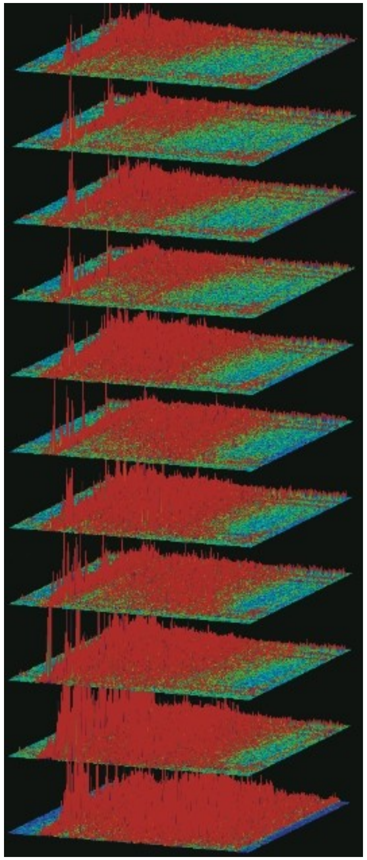


SCX
(Strong Cation Exchange)

RP
(Reverse Phase-C18)

ESI

Ion Trap



– multi-dimensional dataset registration

PTM's
Post-translational Modifications

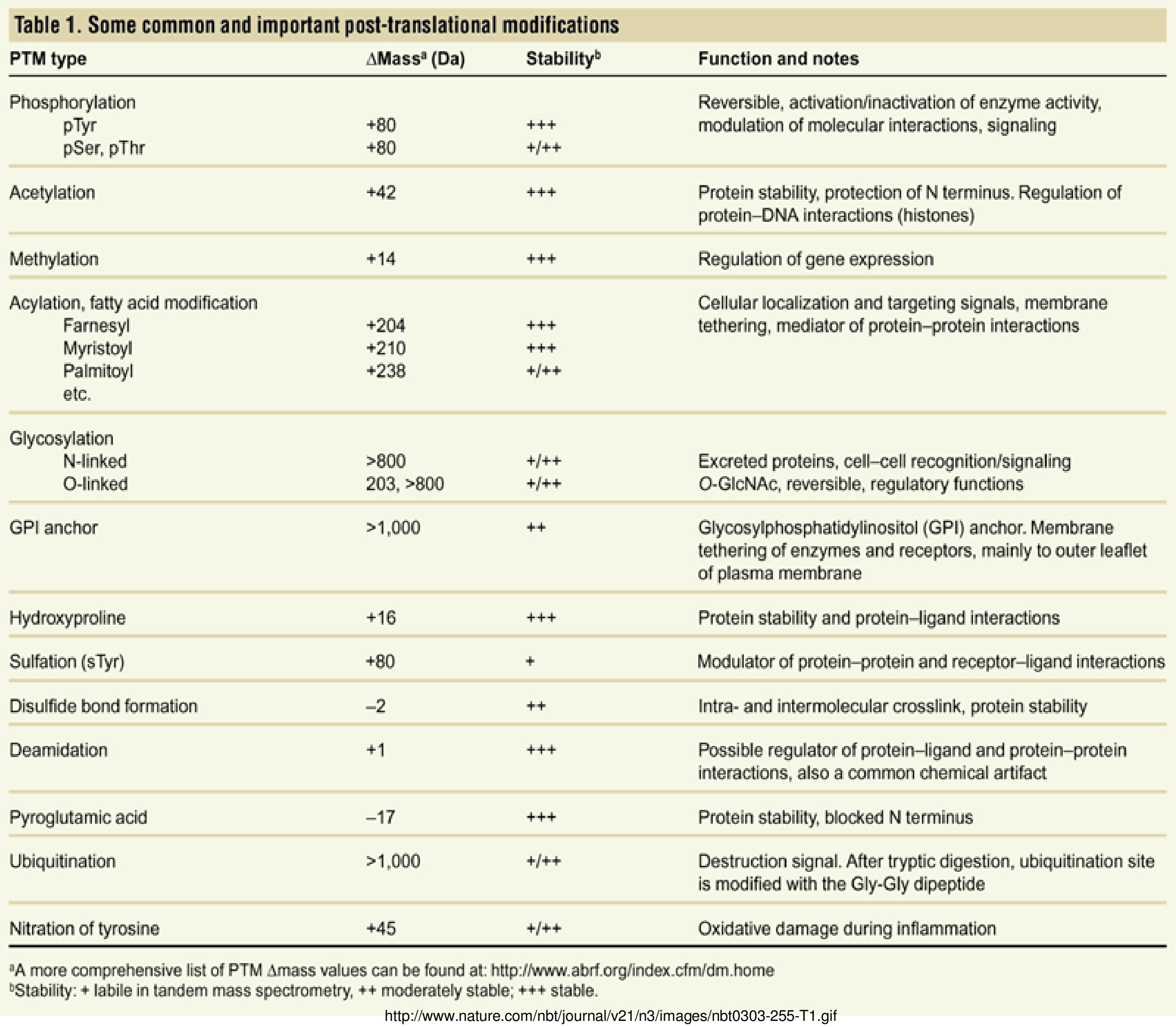


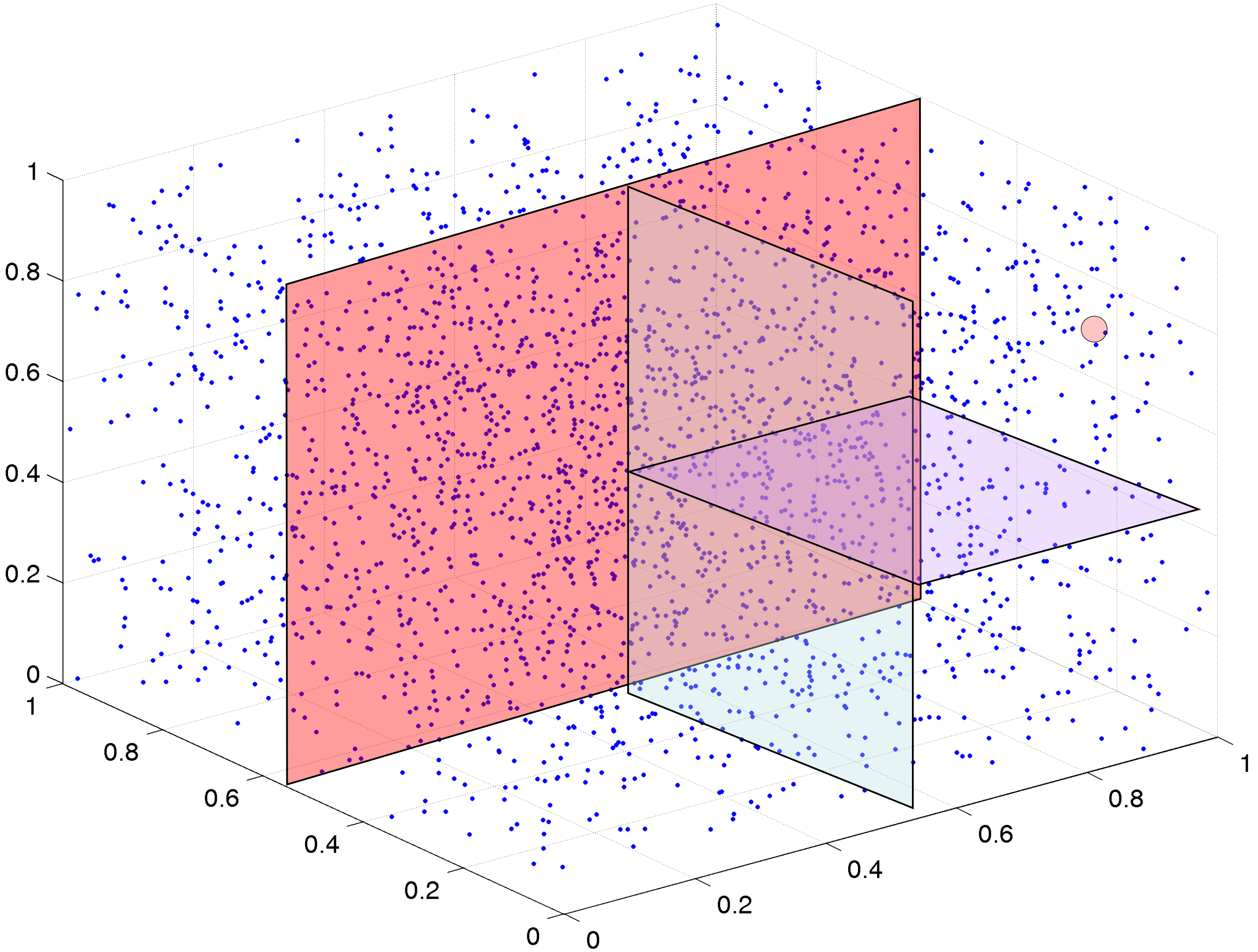
Table 1. Some common and important post-translational modifications

PTM type	ΔMass[a] (Da)	Stability[b]	Function and notes
Phosphorylation			Reversible, activation/inactivation of enzyme activity, modulation of molecular interactions, signaling
pTyr	+80	+++	
pSer, pThr	+80	+/++	
Acetylation	+42	+++	Protein stability, protection of N terminus. Regulation of protein–DNA interactions (histones)
Methylation	+14	+++	Regulation of gene expression
Acylation, fatty acid modification			Cellular localization and targeting signals, membrane tethering, mediator of protein–protein interactions
Farnesyl	+204	+++	
Myristoyl	+210	+++	
Palmitoyl	+238	+/++	
etc.			
Glycosylation			
N-linked	>800	+/++	Excreted proteins, cell–cell recognition/signaling
O-linked	203, >800	+/++	*O*-GlcNAc, reversible, regulatory functions
GPI anchor	>1,000	++	Glycosylphosphatidylinositol (GPI) anchor. Membrane tethering of enzymes and receptors, mainly to outer leaflet of plasma membrane
Hydroxyproline	+16	+++	Protein stability and protein–ligand interactions
Sulfation (sTyr)	+80	+	Modulator of protein–protein and receptor–ligand interactions
Disulfide bond formation	–2	++	Intra- and intermolecular crosslink, protein stability
Deamidation	+1	+++	Possible regulator of protein–ligand and protein–protein interactions, also a common chemical artifact
Pyroglutamic acid	–17	+++	Protein stability, blocked N terminus
Ubiquitination	>1,000	+/++	Destruction signal. After tryptic digestion, ubiquitination site is modified with the Gly-Gly dipeptide
Nitration of tyrosine	+45	+/++	Oxidative damage during inflammation

[a] A more comprehensive list of PTM Δmass values can be found at: http://www.abrf.org/index.cfm/dm.home
[b] Stability: + labile in tandem mass spectrometry, ++ moderately stable; +++ stable.

http://www.nature.com/nbt/journal/v21/n3/images/nbt0303-255-T1.gif

Spectra In Metric-Space

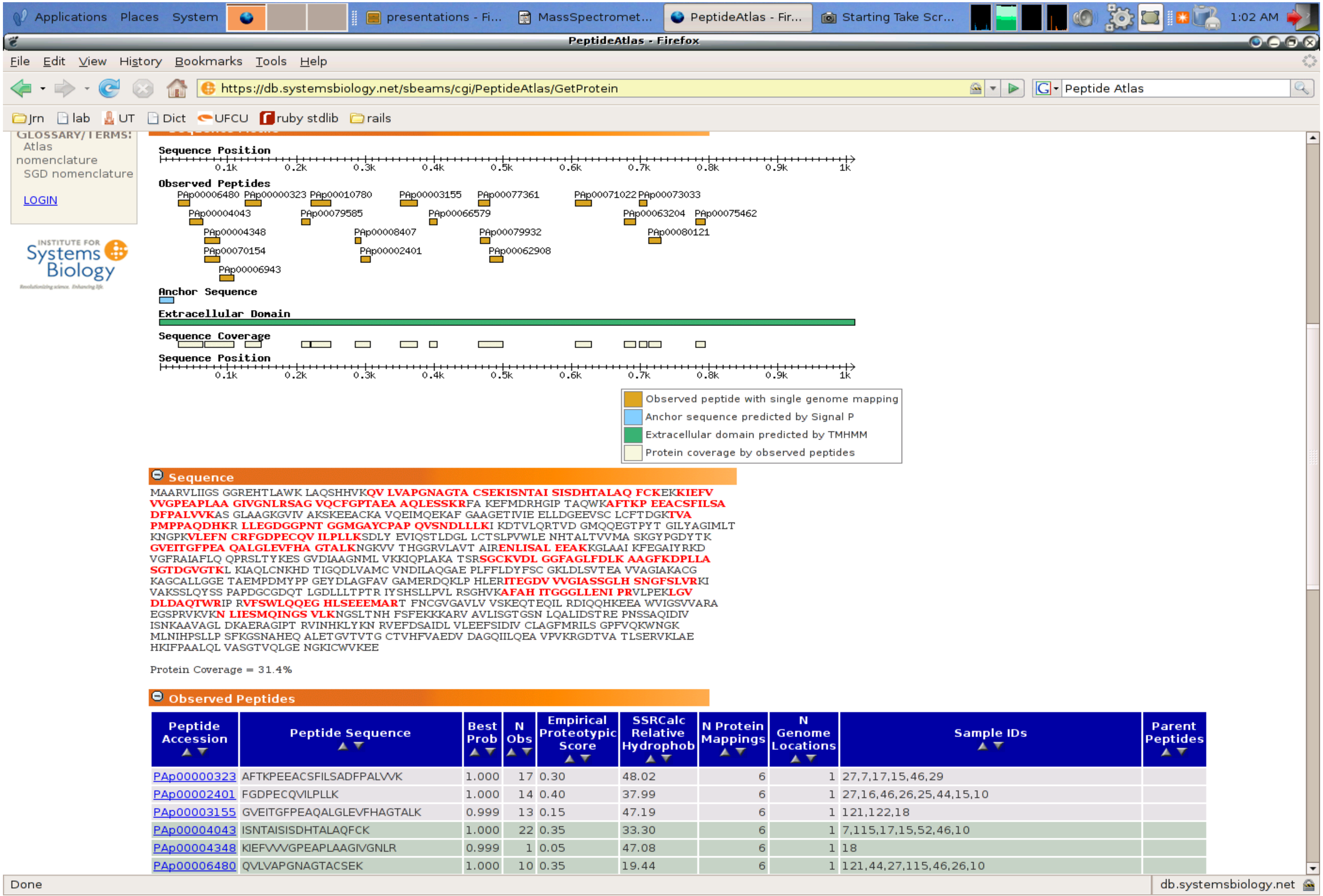


2300 points (at random) in 3D space

Data Format/Storage/Sharing

- Object Models still being worked out
- Huge Datasets
  - how much to save?
  - how much is it worth?
- Sharing
  - OPD
  - Peptide Atlas
  - PRIDE
  - GPM

Biological Integration


Applications Places System
presentations - Fi...
MassSpectromet...
PeptideAtlas - Fir...
Starting Take Scr...
1:02 AM
PeptideAtlas - Firefox
File Edit View History Bookmarks Tools Help
https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetProtein
Peptide Atlas
Jrn lab UT Dict UFCU ruby stdlib rails
GLOSSARY/TERMS:
Atlas nomenclature
SGD nomenclature
LOGIN
INSTITUTE FOR Systems Biology
Sequence Position
0.1k 0.2k 0.3k 0.4k 0.5k 0.6k 0.7k 0.8k 0.9k 1k
Observed Peptides
PAp00006480 PAp00000323 PAp00010780 PAp00003155 PAp00077361 PAp00071022 PAp00073033
PAp00004043 PAp00079585 PAp00066579 PAp00063204 PAp00075462
PAp00004348 PAp00008407 PAp00079932 PAp00080121
PAp00070154 PAp00002401 PAp00062908
PAp00006943
Anchor Sequence
Extracellular Domain
Sequence Coverage
Sequence Position
0.1k 0.2k 0.3k 0.4k 0.5k 0.6k 0.7k 0.8k 0.9k 1k
Observed peptide with single genome mapping
Anchor sequence predicted by Signal P
Extracellular domain predicted by TMHMM
Protein coverage by observed peptides
Done
db.systemsbiology.net

Sequence

MAARVLIIGS GGREHTLAWK LAQSHHVK**QV LVAPGNAGTA CSEKISNTAI SISDHTALAQ FCK**EK**KIEFV**
VVGPEAPLAA GIVGNLRSAG VQCFGPTAEA AQLESSKRFA KEFMDRHGIP TAQWK**AFTKP EEACSFILSA**
DFPALVVKAS GLAAGKGVIV AKSKEEACKA VQEIMQEKAF GAAGETIVIE ELLDGEEVSC LCFTDGK**TVA**
PMPPAQDHKR **LLEGDGGPNT GGMGAYCPAP QVSNDLLLK**I KDTVLQRTVD GMQQEGTPYT GILYAGIMLT
KNGPK**VLEFN CRFGDPECQV ILPLLK**SDLY EVIQSTLDGL LCTSLPVWLE NHTALTVVMA SKGYPGDYTK
GVEITGFPEA QALGLEVFHA GTALKNGKVV THGGRVLAVT AIR**ENLISAL EEAK**KGLAAI KFEGAIYRKD
VGFRAIAFLQ QPRSLTYKES GVDIAAGNML VKKIQPLAKA TSR**SGCKVDL GGFAGLFDLK AAGFKDPLLA**
SGTDGVGTKL KIAQLCNKHD TIGQDLVAMC VNDILAQGAE PLFFLDYFSC GKLDLSVTEA VVAGIAKACG
KAGCALLGGE TAEMPDMYPP GEYDLAGFAV GAMERDQKLP HLER**ITEGDV VVGIASSGLH SNGFSLVR**KI
VAKSSLQYSS PAPDGCGDQT LGDLLLTPTR IYSHSLLPVL RSGHVK**AFAH ITGGGLLENI PR**VLPEK**LGV**
DLDAQTWRIP R**VFSWLQQEG HLSEEEMAR**T FNCGVGAVLV VSKEQTEQIL RDIQQHKEEA WVIGSVVARA
EGSPRVKVK**N LIESMQINGS VLK**NGSLTNH FSFEKKARV AVLISGTGSN LQALIDSTRE PNSSAQIDIV
ISNKAAVAGL DKAERAGIPT RVINHKLYKN RVEFDSAIDL VLEEFSIDIV CLAGFMRILS GPFVQKWNGK
MLNIHPSLLP SFKGSNAHEQ ALETGVTVTG CTVHFVAEDV DAGQIILQEA VPVKRGDTVA TLSERVKLAE
HKIFPAALQL VASGTVQLGE NGKICWVKEE

Protein Coverage = 31.4%

Observed Peptides

Peptide Accession ▲▼	Peptide Sequence ▲▼	Best Prob ▲▼	N Obs ▲▼	Empirical Proteotypic Score ▲▼	SSRCalc Relative Hydrophob ▲▼	N Protein Mappings ▲▼	N Genome Locations ▲▼	Sample IDs ▲▼	Parent Peptides ▲▼
PAp00000323	AFTKPEEACSFILSADFPALVVK	1.000	17	0.30	48.02	6	1	27,7,17,15,46,29	
PAp00002401	FGDPECQVILPLLK	1.000	14	0.40	37.99	6	1	27,16,46,26,25,44,15,10	
PAp00003155	GVEITGFPEAQALGLEVFHAGTALK	0.999	13	0.15	47.19	6	1	121,122,18	
PAp00004043	ISNTAISISDHTALAQFCK	1.000	22	0.35	33.30	6	1	7,115,17,15,52,46,10	
PAp00004348	KIEFVVVGPEAPLAAGIVGNLR	0.999	1	0.05	47.08	6	1	18	
PAp00006480	QVLVAPGNAGTACSEK	1.000	10	0.35	19.44	6	1	121,44,27,115,46,26,10	

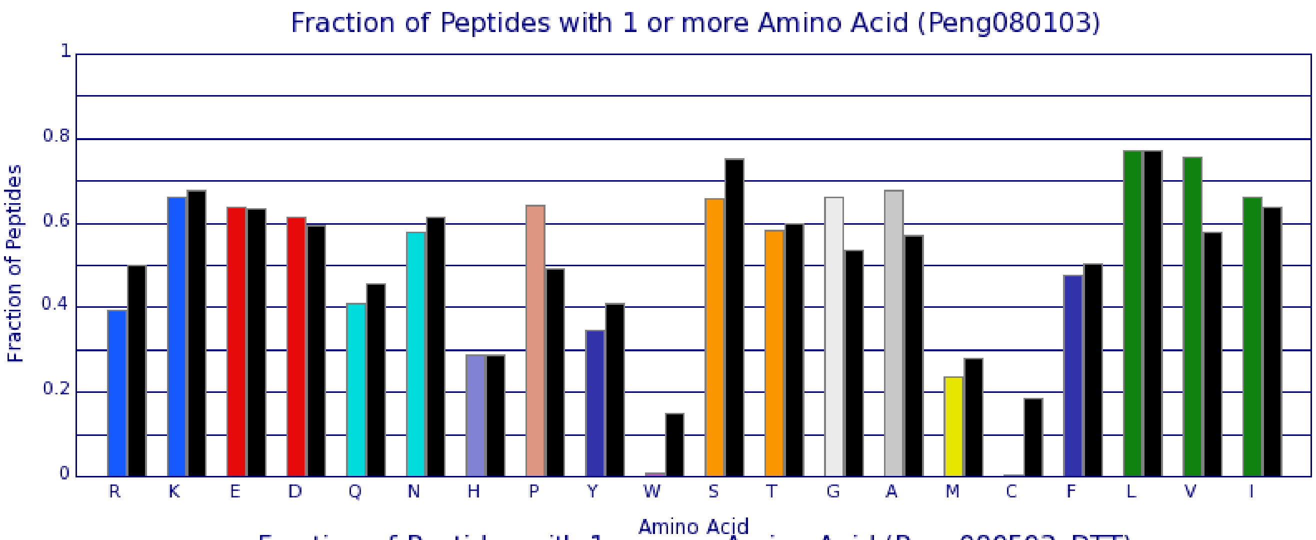
mRNA vs. Protein

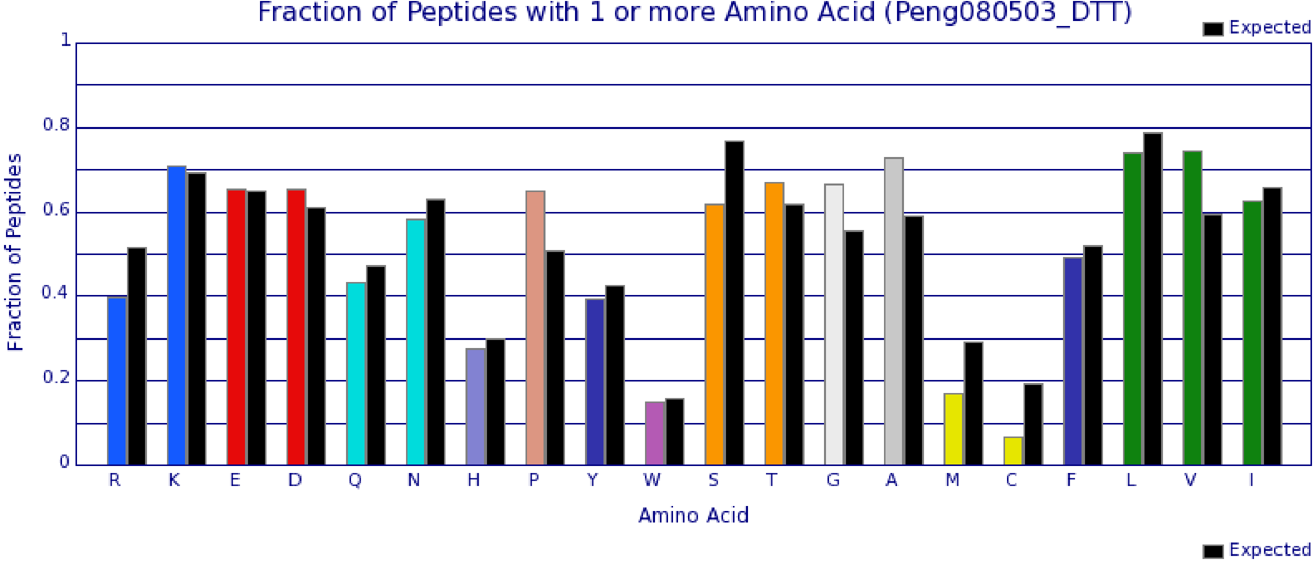
Source	Subject	Perturbation (or sample)	Num Genes	Correlation
Ideker et. al.	Yeast	+/- gal (gal inducing media)	289	$r_p = 0.61$
Futcher et. al.	Yeast	2% ethanol/ 2% glucose	148	$r_s = 0.74$ $r_p = 0.76$ [a]
Washburn et. al.	Yeast	rich/minimal	678	$r_s = 0.45$
Griffin et. al.	Yeast	2% ethanol/ 2% galactose	245	$r_s = 0.21$
Gygi et. al.	Yeast	mid-log	106	$r_p = 0.94$ $r_s = 0.59$ [b] $r_p = 0.356$ [c]
Chen et. al.	Lung adenocarcinomas	57 stage I, 19 stage III, 9 non-neoplastic	98 (165 prots)	$r_p = -0.025$ [d]

a = after normalizing the data
b = calculated by Futcher et. al.
c = 73 genes with lower abundance transcripts
d = after detailed statistical analysis

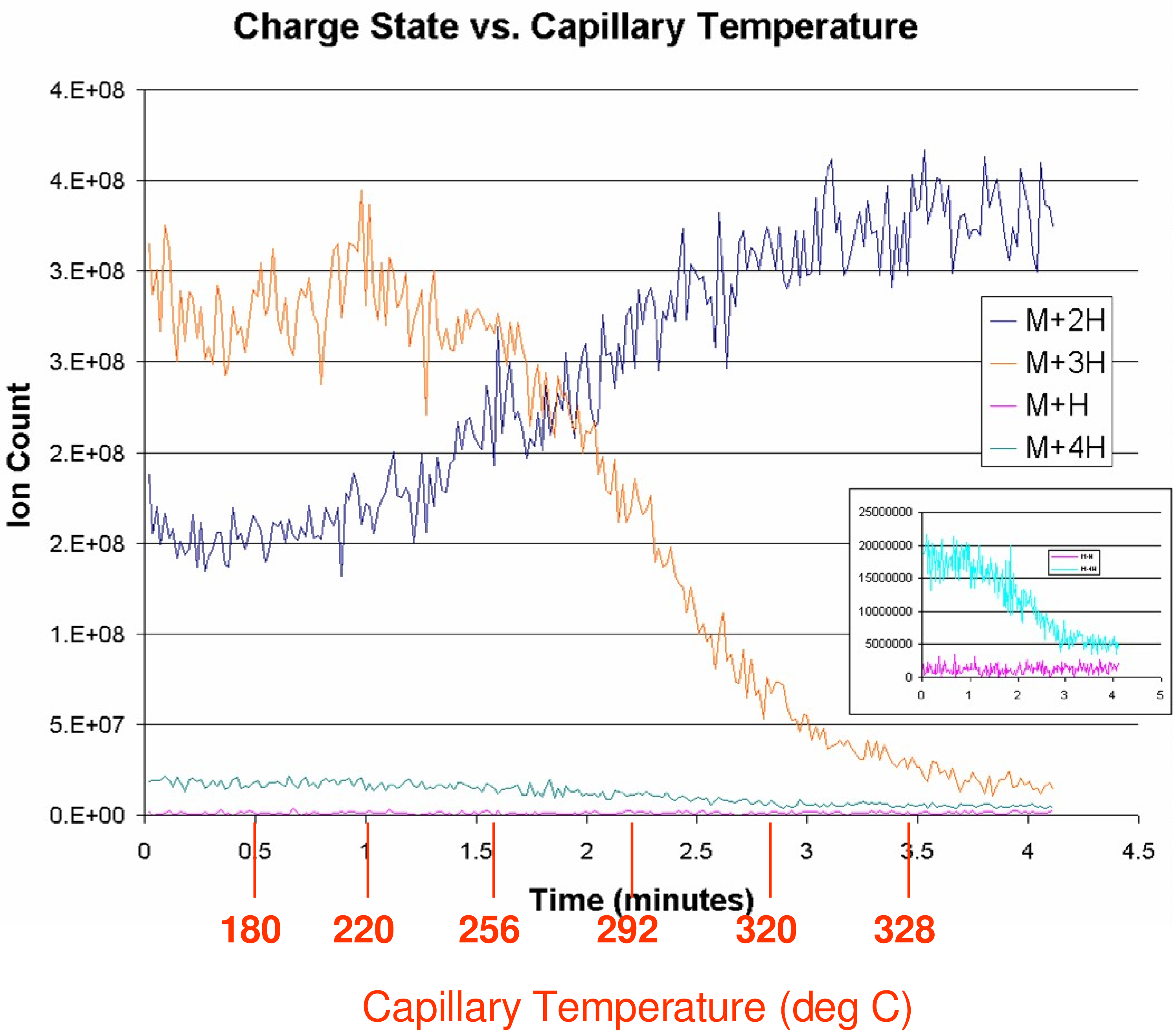
Disulfide Bonds

Expected fraction: 1 - (1-freq)^n
[Rolling one “6” in **n** rolls is 1 – (5/6)^n]





Using RasMol amino acid color scheme


Charge State vs. Capillary Temperature
4.E+08
4.E+08
3.E+08
3.E+08
2.E+08
2.E+08
1.E+08
5.E+07
0.E+00
Ion Count
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Time (minutes)
M+2H
M+3H
M+H
M+4H
25000000
20000000
15000000
10000000
5000000
0
0
1
2
3
4
5
M+H
M+4H
180
220
256
292
320
328
Capillary Temperature (deg C)

Acknowledgments

- Dr. Edward Marcotte
- Dr. Klaus Linse
- Dr. Maria Person
- Dr. Aleksey Nakorshevskiy
- Dr. Rong Wang
- Dr. Peng Lu
- Zhihua Li