

# Working at the Interface between Proteomics & Informatics at Indiana University

Randy J. Arnold

Proteomics Manager

National Center for Glycomics & Glycoproteomics

I533

March 23, 2006

# Overview

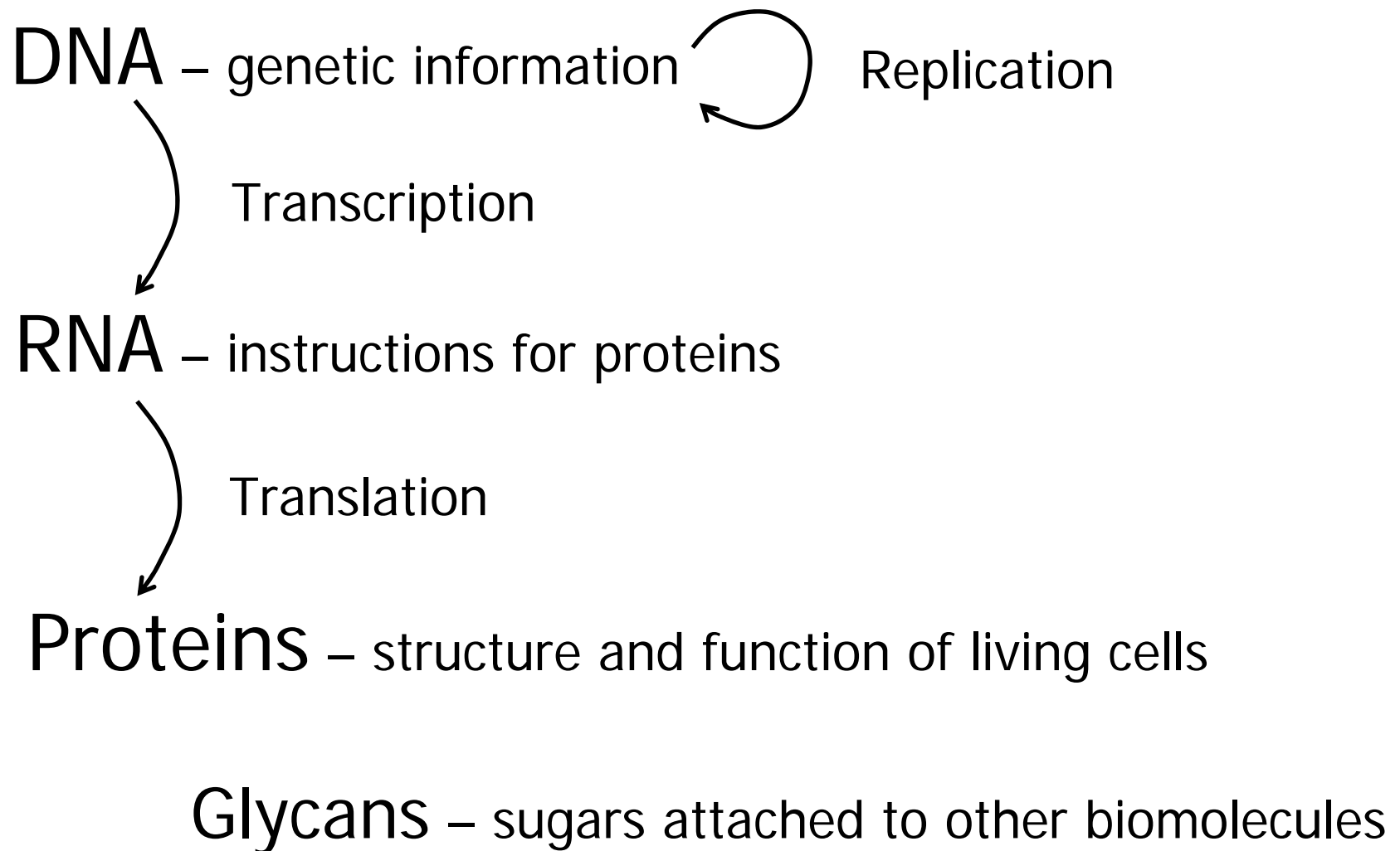
- Introduction to Proteomics
- Proteomics Platforms / Approaches
- Peptide Fragmentation – Predictable?
- Biology-based Separations
- Influence of Informatics

# What is Proteomics?

Proteomics is the study of protein expression, regulation, modification, and function in living systems for understanding how living systems use proteins. Using a variety of techniques, proteomics can be used to study how proteins interact within a system, or how proteins change due to applied stresses.

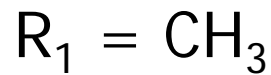
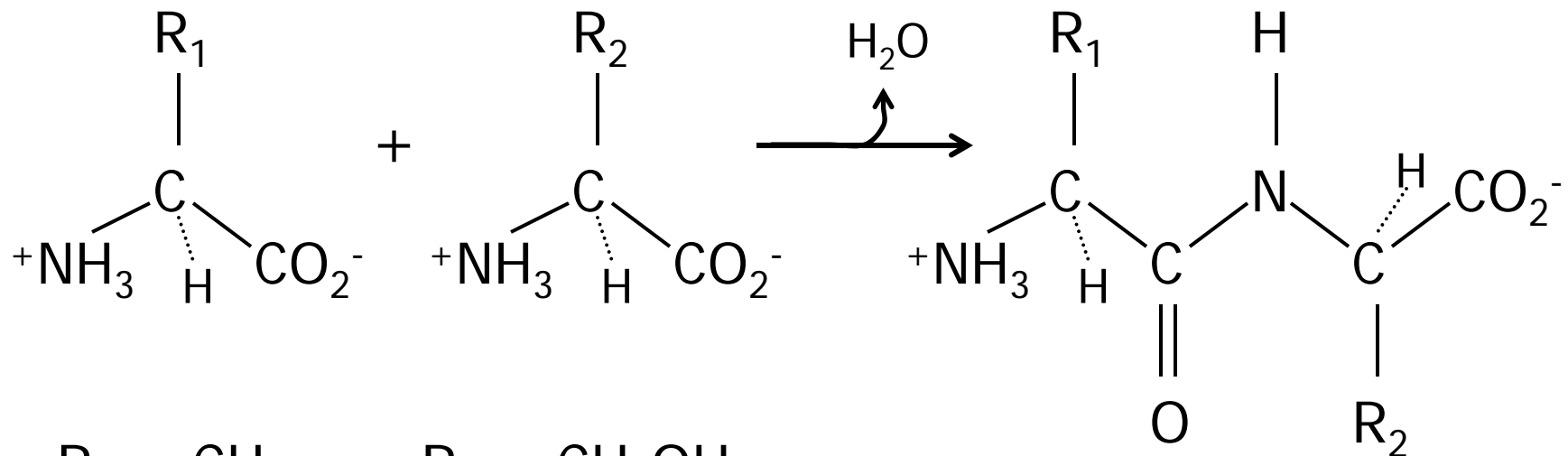
Proteomics requires the use of advanced measurement techniques with an emphasis on separations and mass spectrometry.

# The Central Dogma of Life



# The Basics - Proteins

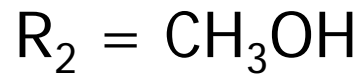
Polymers of amino acids (20 naturally occurring)



Alanine

Ala

A



Serine

Ser

S

Alanylserine

Ala-Ser

AS

# The Basics – Protein Structure

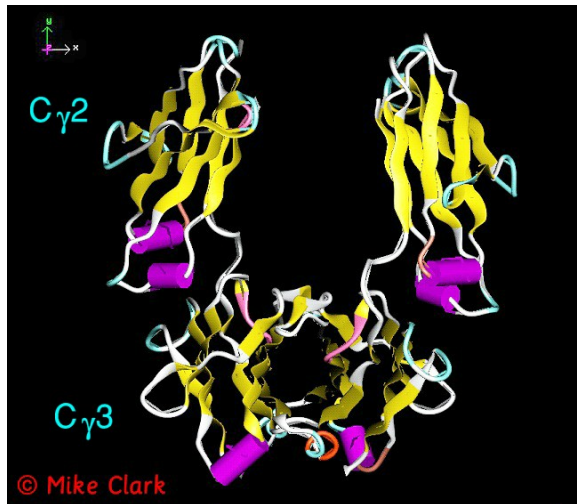
primary structure – amino acid sequence

DRLEFIVTALLKPW

N-terminus

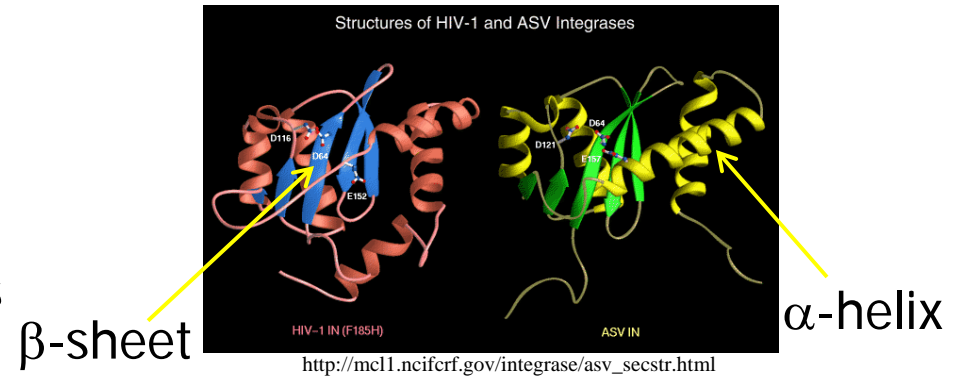
C-terminus

tertiary structure – protein folding

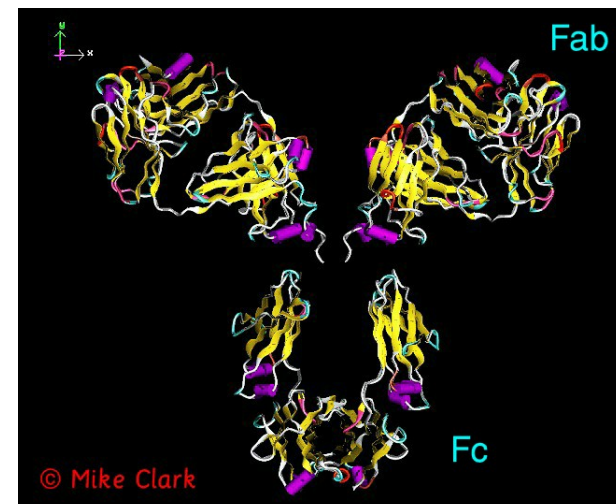


<http://www.path.cam.ac.uk/~mrc7/igs/mikeimages.html>

secondary structure – local spatial arrangement



quaternary structure – multimeric complexes

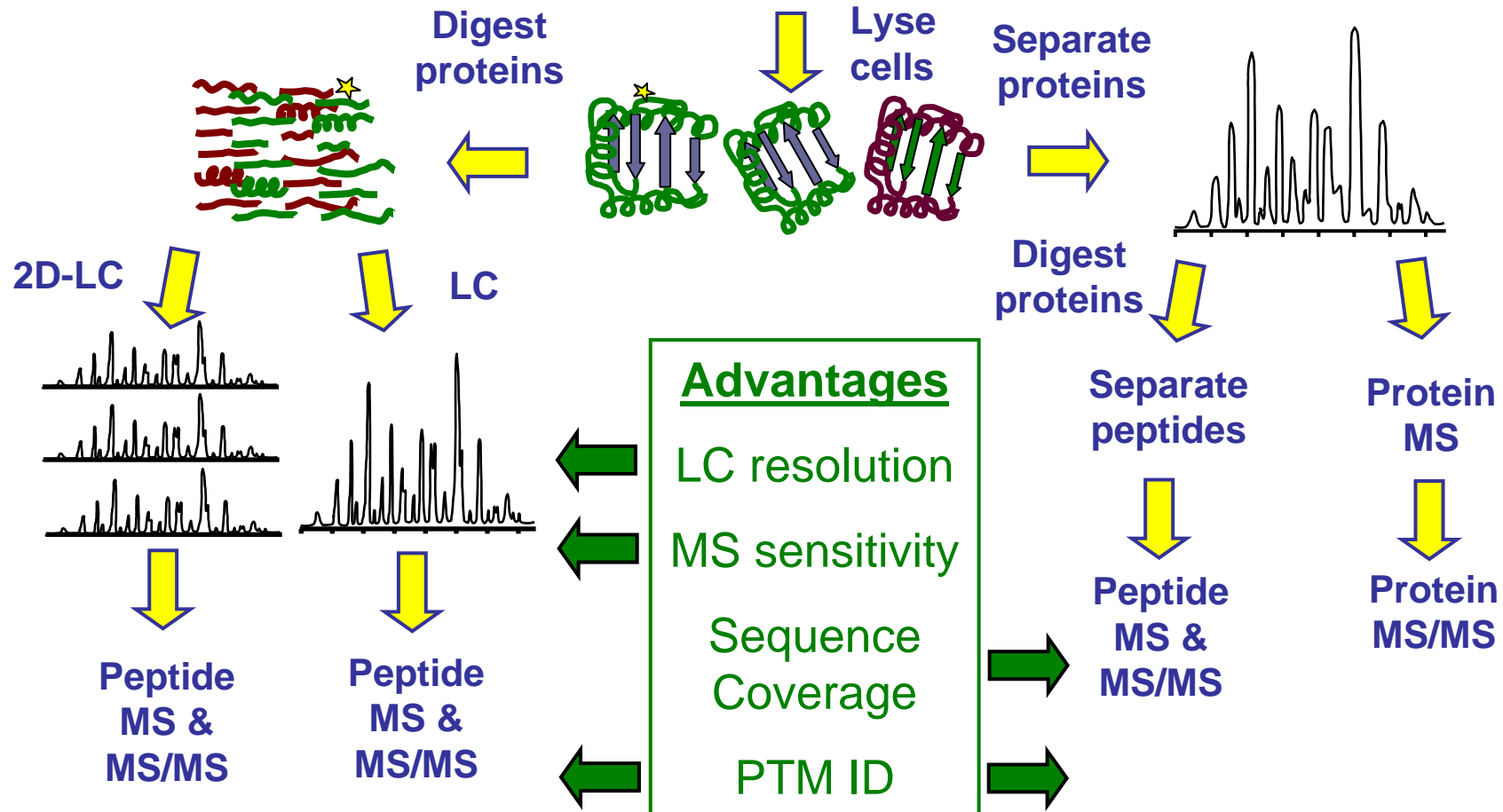
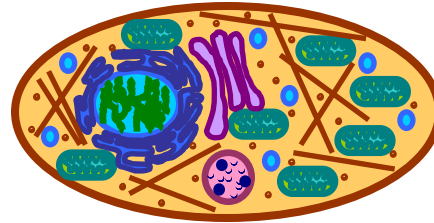


<http://www.path.cam.ac.uk/~mrc7/igs/mikeimages.html>

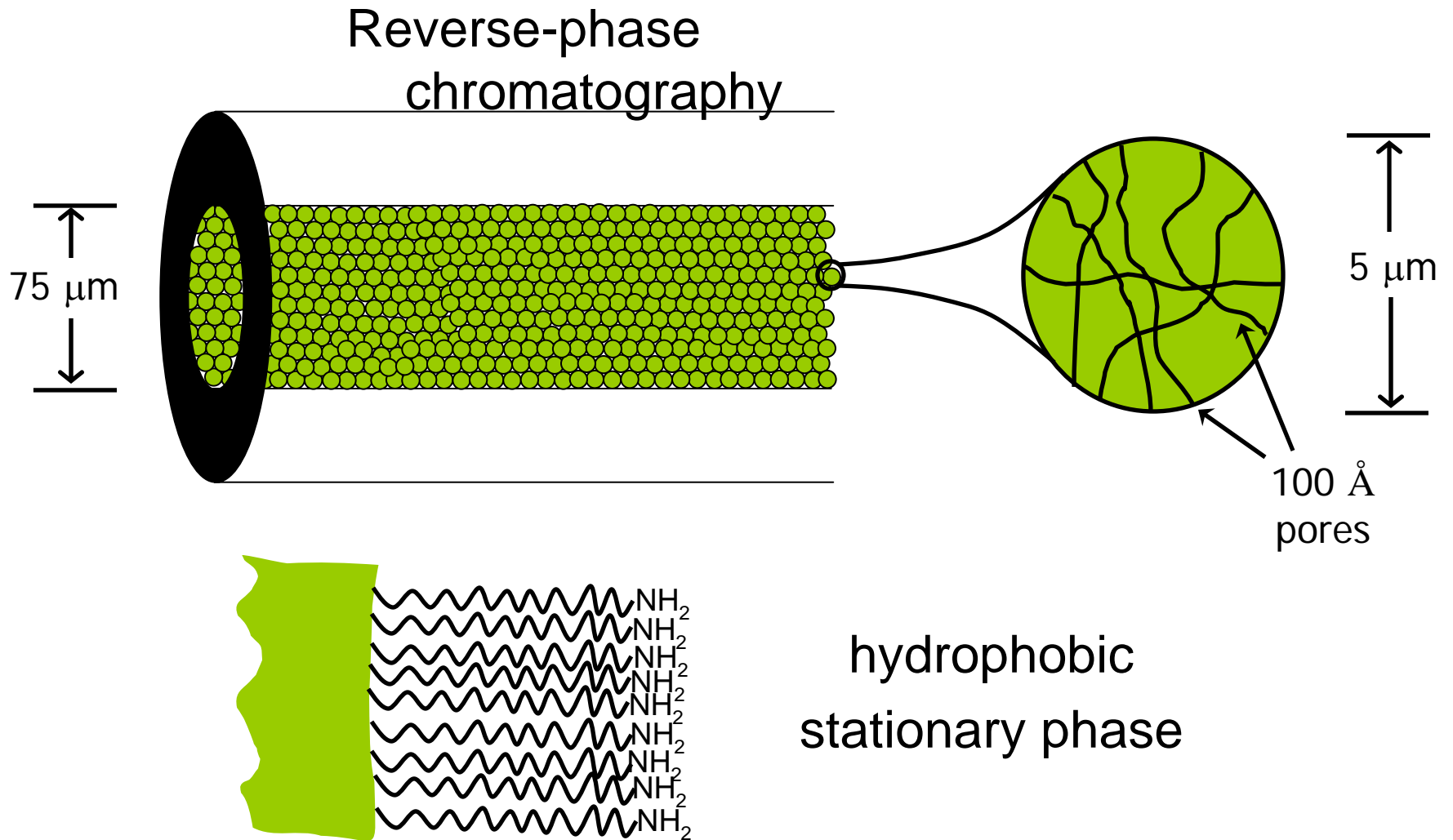
# Proteomics Approaches

## Bottom-up

## Top-down



# Chromatographic Separation

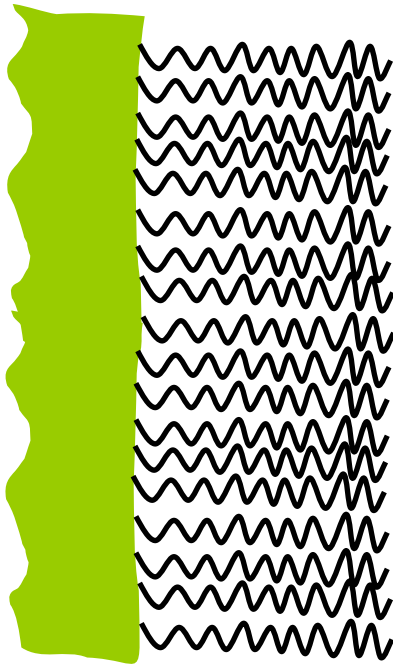


# Chromatographic Separation

## Molecular Interactions

+NPLTGLAK+

+MILSGVAR+



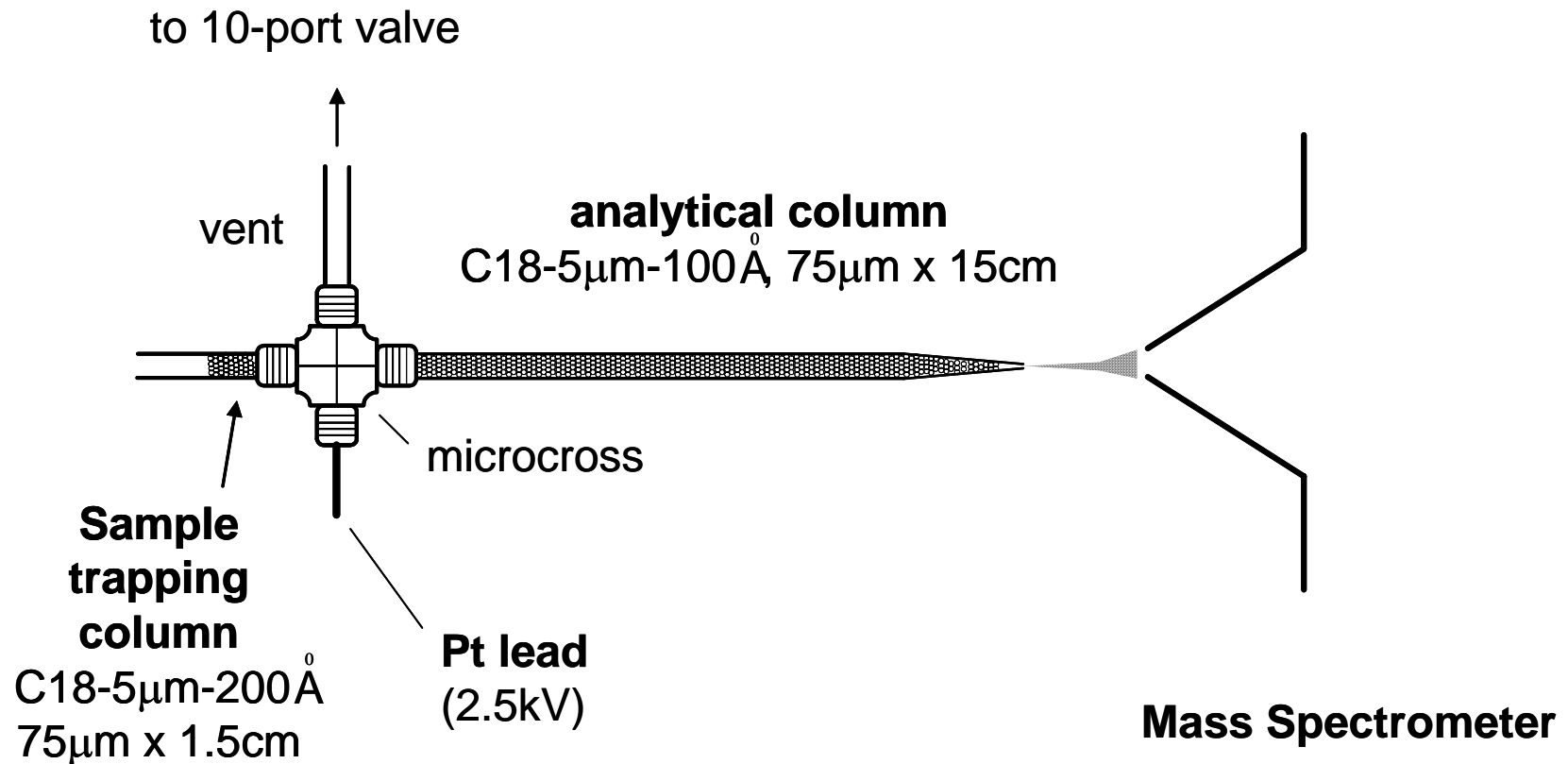
1) 95% aqueous / 5% organic mobile phase

2) 75% aqueous / 25% organic mobile phase

3) 65% aqueous / 35% organic mobile phase

# Electrospray Ionization

2002 Nobel Prize in Chemistry awarded to John Fenn for advancing electrospray



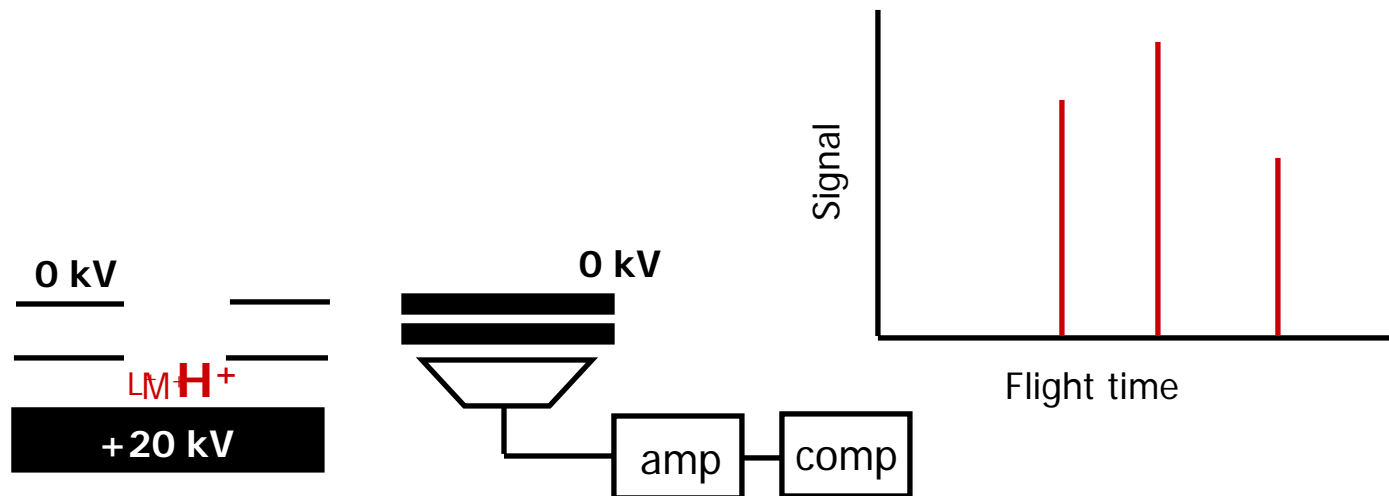
Compliments of Dr. Myeong Hee Moon

# Mass Spectrometry

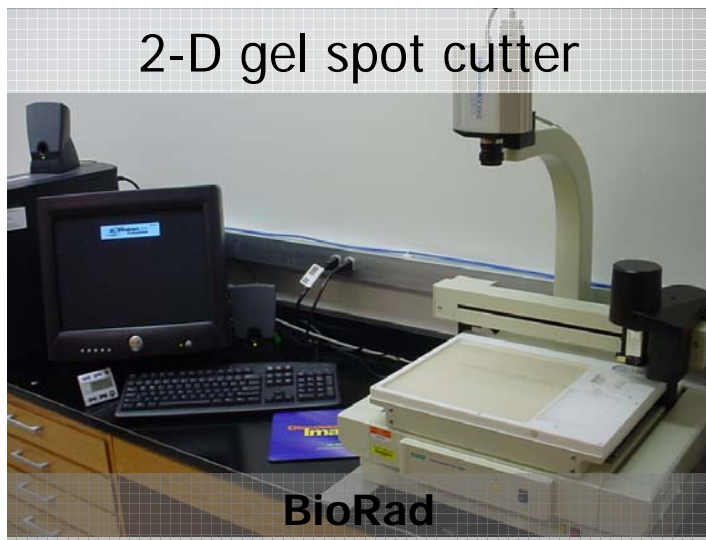
## Time-of-Flight MS



- 1) Ions enter source region, accelerated toward reflectron.
- 2) Ions separate in space based on their relative mass-to-charge ( $m/z$ ).
- 3) Ions reverse path in reflectron.
- 4) Ions impact detector.

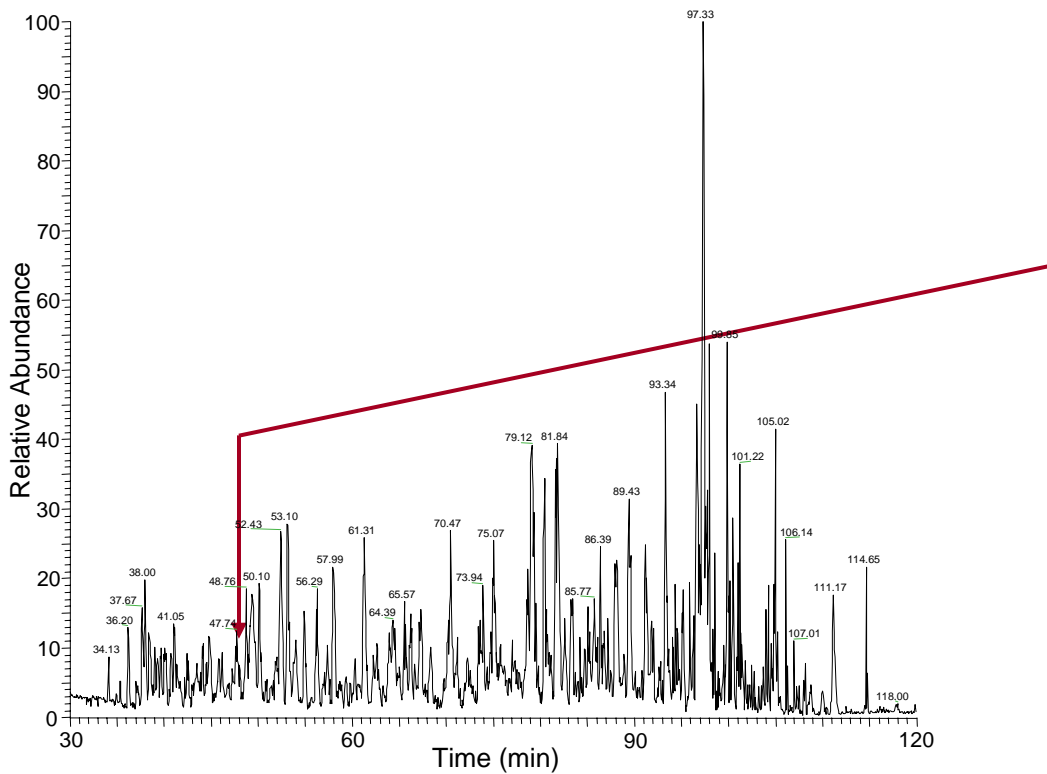


# Instrumentation

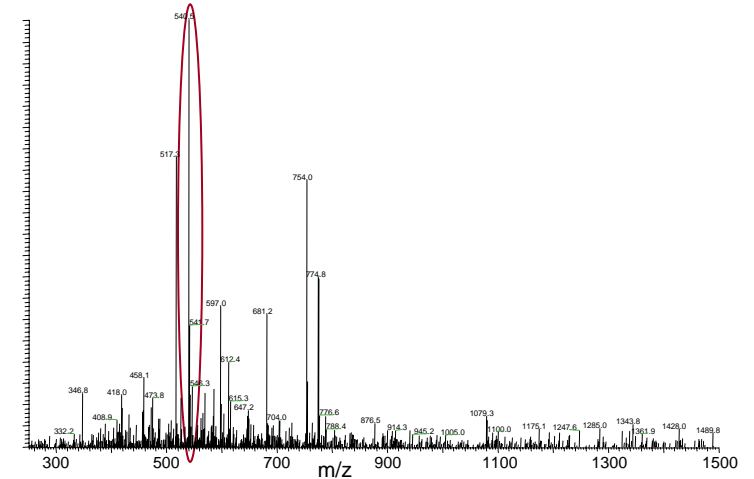


# nanoLC-MS/MS Data

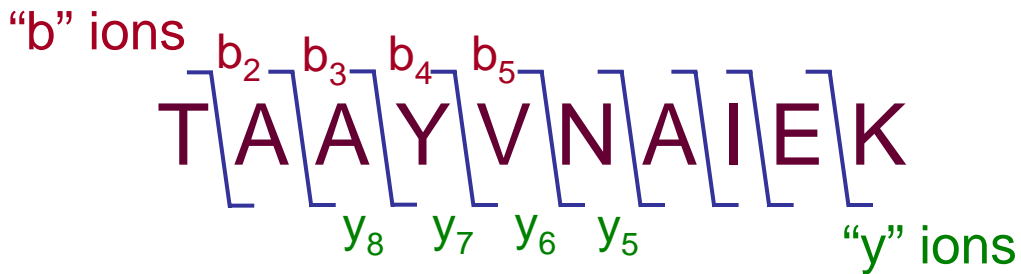
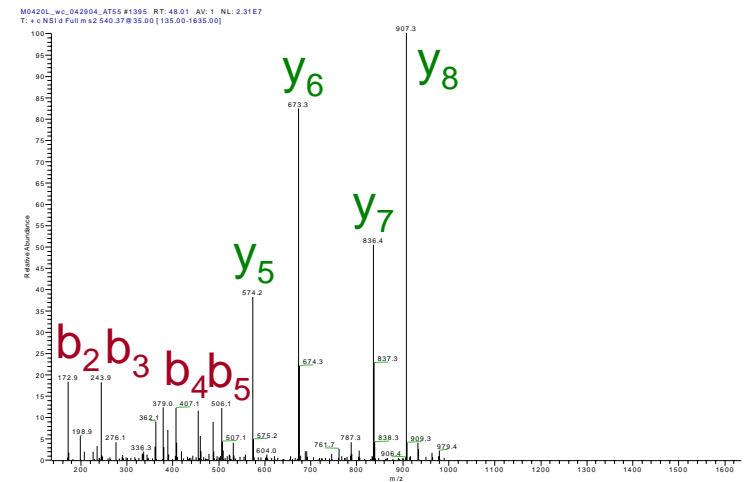
Base Peak Chromatogram



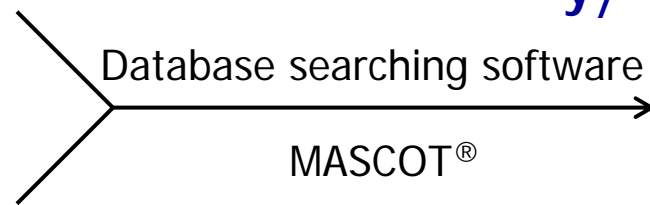
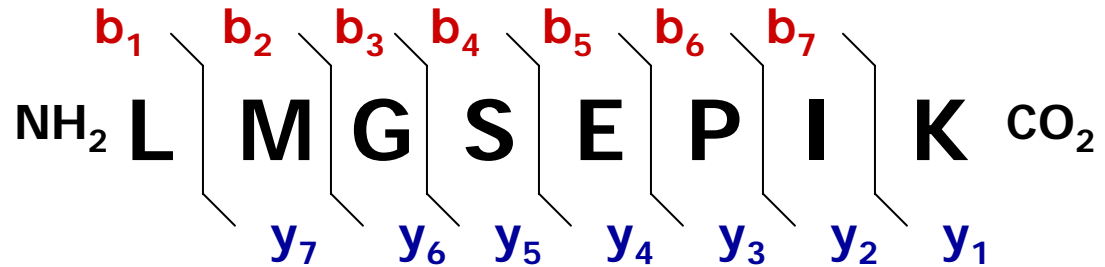
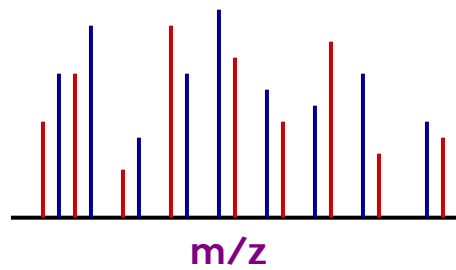
Mass Spectrum at 48.08 min.



Tandem MS of m/z 540.4



# Database Searching (Informatics)



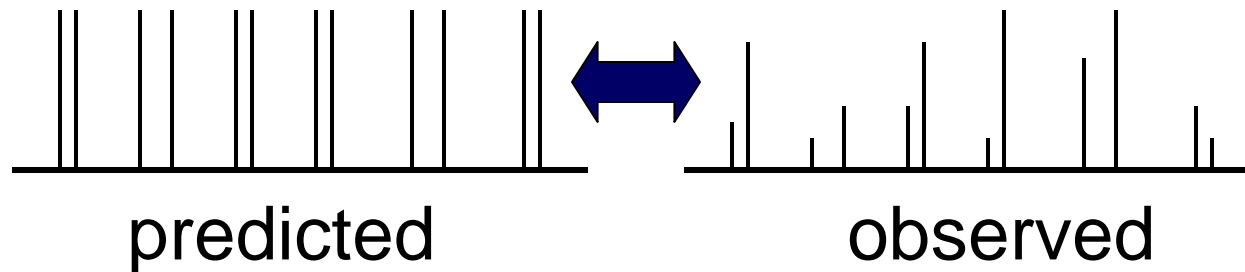
Results

Database (SwissProt)	
Actin	MYTCVPIASEQUENCEMIMIEWTPOS DLIRPTVCIMNERCVGGPYILCMTEND
Amylase	DSLIKRNYTIPMCSQIRECNHIPLMTRCH GYYKWSIALAINTQSFGIVRIVAMNKLPS SCRTIVGHWEDRICTMQNCISPPEKELIA VARGTSP
...	

Proteins found			
Hemoglobin, beta chain			
Pept.	Mass	Score	Sequence
1	738.84	41	HLDNLK
2	912.01	61	VHLTDAEK
3	915.06	56	AAVNGLWGK
4	1090.24	41	VINAFNDGLK
5	1122.33	62	VVAGVASALAHK
6	1218.42	70	LVINAFNDGLK
...			

# Identification approaches

- SEQUEST<sup>1</sup> – cross-correlation



- Mascot<sup>2</sup> – probability-based scoring

–  $b_1$   $b_2$   **$b_3$**   **$b_4$**   **$b_5$**   $b_6$   **$b_7$**   $b_8$   **$b_9$**   **$b_{10}$**   $b_{11}$

–  $y_{11}$   $y_{10}$   **$y_9$**   **$y_8$**   **$y_7$**   **$y_6$**   **$y_5$**   $y_4$   $y_3$   **$y_2$**   $y_1$

1. Yates JR, III, Eng JK, McCormack AL, Schieltz D. *Anal Chem.* **67** 1426 (1995).
2. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. *Electrophoresis.* **20** 3551 (1999).

# Intensity Prediction

- Kinetic model<sup>1</sup>
  - model based on chemical mechanism (CID)
  - universally applicable?
- Decision tree<sup>2</sup>
  - +2 only
  - b & y ions only

1. Zhang Z. *Anal Chem.* **76** 3908 (2004).

2. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. *Nat Biotechnol.* **22** 214 (2004).

# **A Machine Learning Approach to Predicting Peptide Fragmentation Spectra**

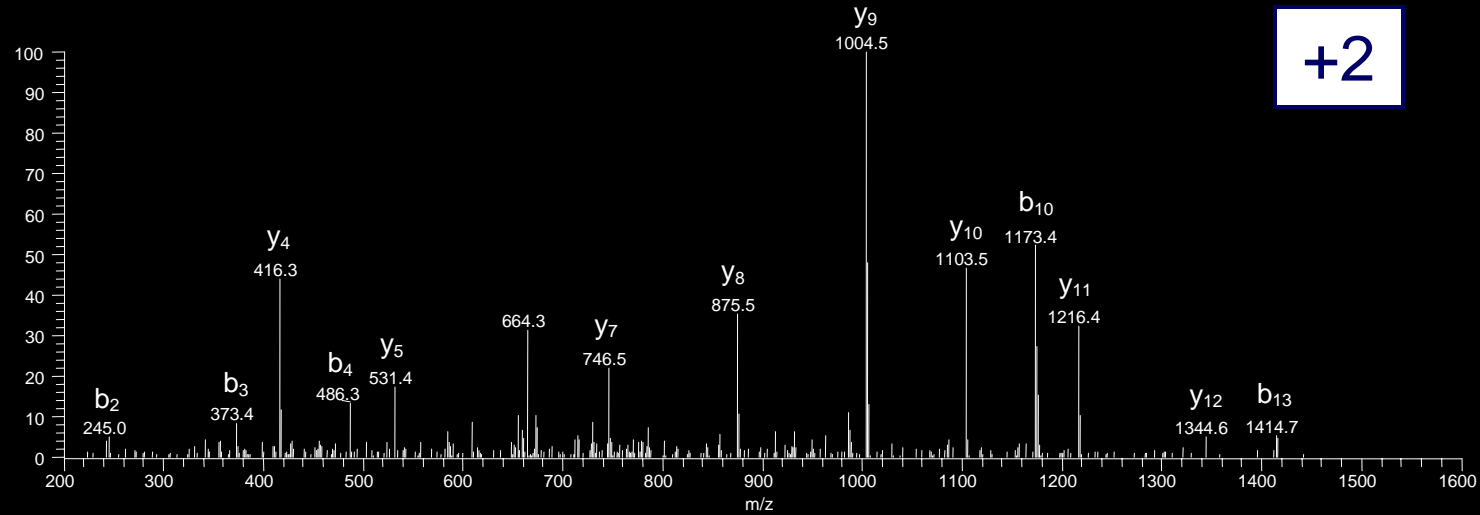
Randy J. Arnold, Narmada Jayasankar, Divya Aggarwal, Haixu Tang, and Predrag Radivojac

**Collaborative project between the Department of  
Chemistry and the School of Informatics**

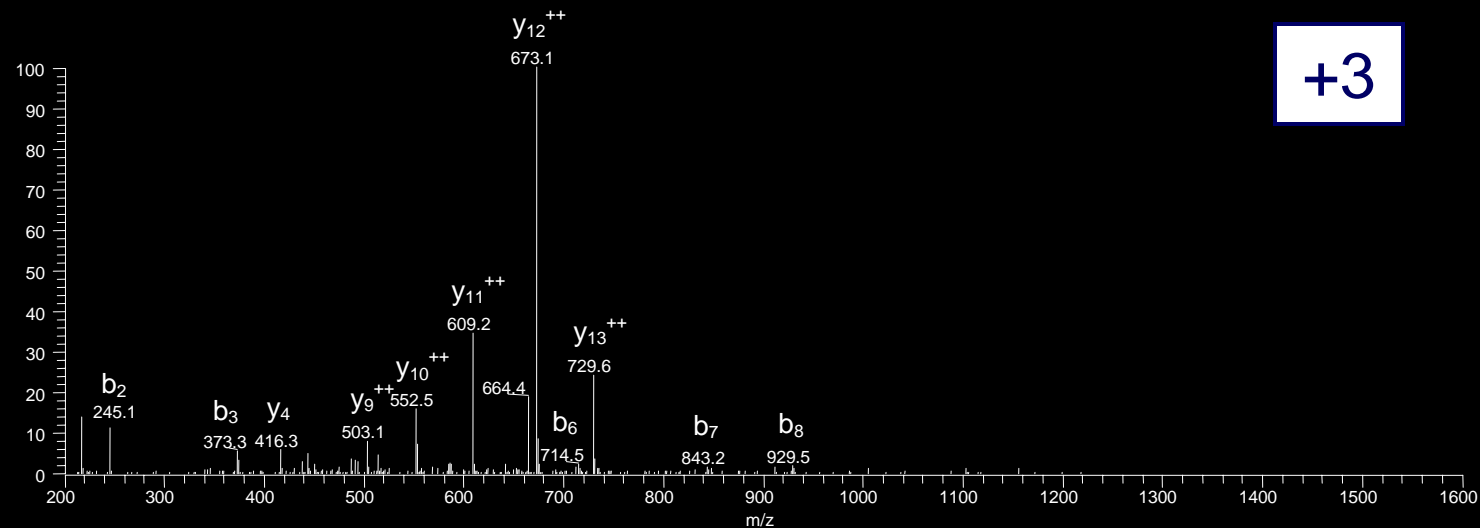
**Presented at PSB in Jan. 2006**

# Peptide charge state

MLQLVEESKDAGIR



+2



+3

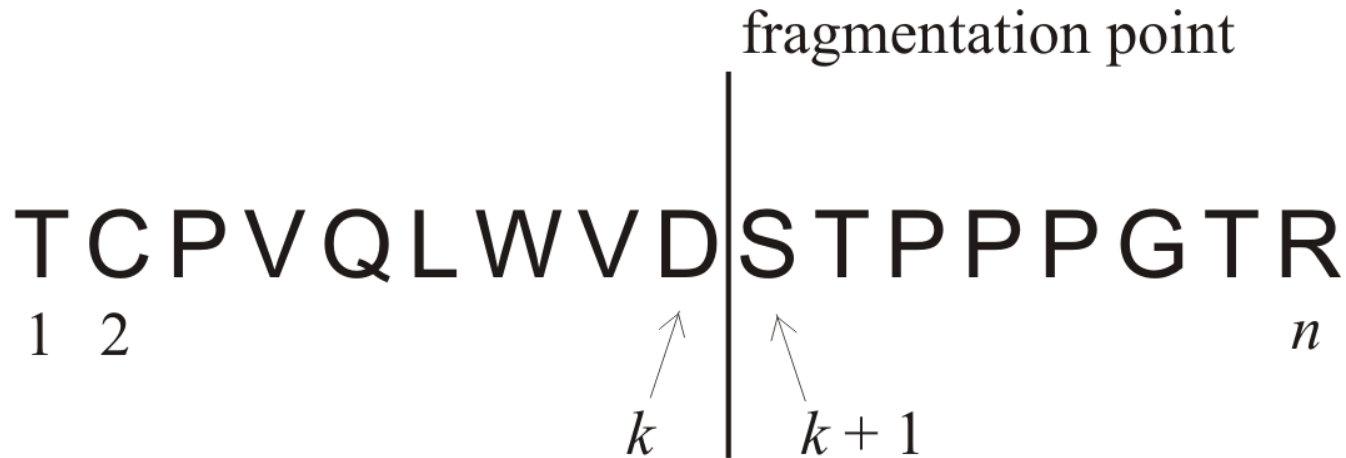
# Method – classification model

- Precursor sequence  $S$ 
  - Charge  $q_S \in \{+2, +3\}$
- Estimate probabilities:  $P(I(i) \geq t \mid S, q_S)$ 
  - where  $I(i)$  is *peak intensity of any fragment ion*:
    - $i \in \{\text{precursor-H}_2\text{O}, b, b\text{-H}_2\text{O}, b\text{-NH}_3, b\text{-H}_2\text{O-NH}_3, y, y\text{-H}_2\text{O}, y\text{-NH}_3, y\text{-H}_2\text{O-NH}_3, b^{2+}, y^{2+}\}$
    - $t = 1\%$  of total intensity of the spectrum

# Datasets

<i>Ion</i>	<i>Doubly charged precursors</i>			<i>Triply charged precursors</i>		
	<i>Positives</i>	<i>Negatives</i>	<i>Total</i>	<i>Positives</i>	<i>Negatives</i>	<i>Total</i>
<i>precursor</i> – H <sub>2</sub> O	239	1484	1723	64	590	654
<i>b</i>	5210	16916	22126	950	12000	12950
<i>b</i> – H <sub>2</sub> O	1700	20426	22126	206	12744	12950
<i>b</i> – NH <sub>3</sub>	678	21448	22126	117	12833	12950
<i>b</i> – H <sub>2</sub> O – NH <sub>3</sub>	249	21877	22126	121	12829	12950
<i>b</i> <sup>2</sup>	-	-	-	1343	11607	12950
<i>y</i>	9323	12802	22126	1639	11311	12950
<i>y</i> – H <sub>2</sub> O	431	21695	22126	132	12818	12950
<i>y</i> – NH <sub>3</sub>	286	21840	22126	101	12849	12950
<i>y</i> – H <sub>2</sub> O – NH <sub>3</sub>	145	21981	22126	107	12843	12950
<i>y</i> <sup>2</sup>	-	-	-	1953	10997	12950

# Features – 202 in total



Tryptic peptide,  $n$  residues long, from human p53

- amino acids at positions  $k$ ,  $k - 1$ ,  $k + 1$ ,  $k + 2$
- amino acid at position 1
- amino acid compositions for both fragment ions
- length and mass of each fragment ion
- various physical / chemical properties

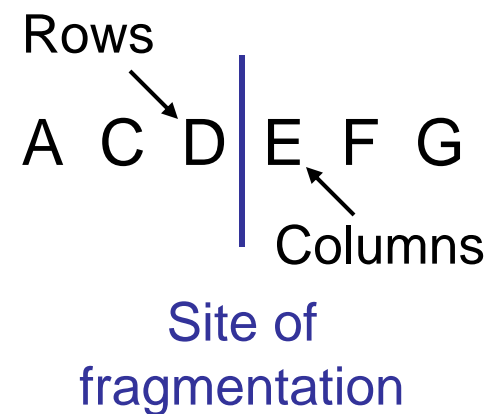
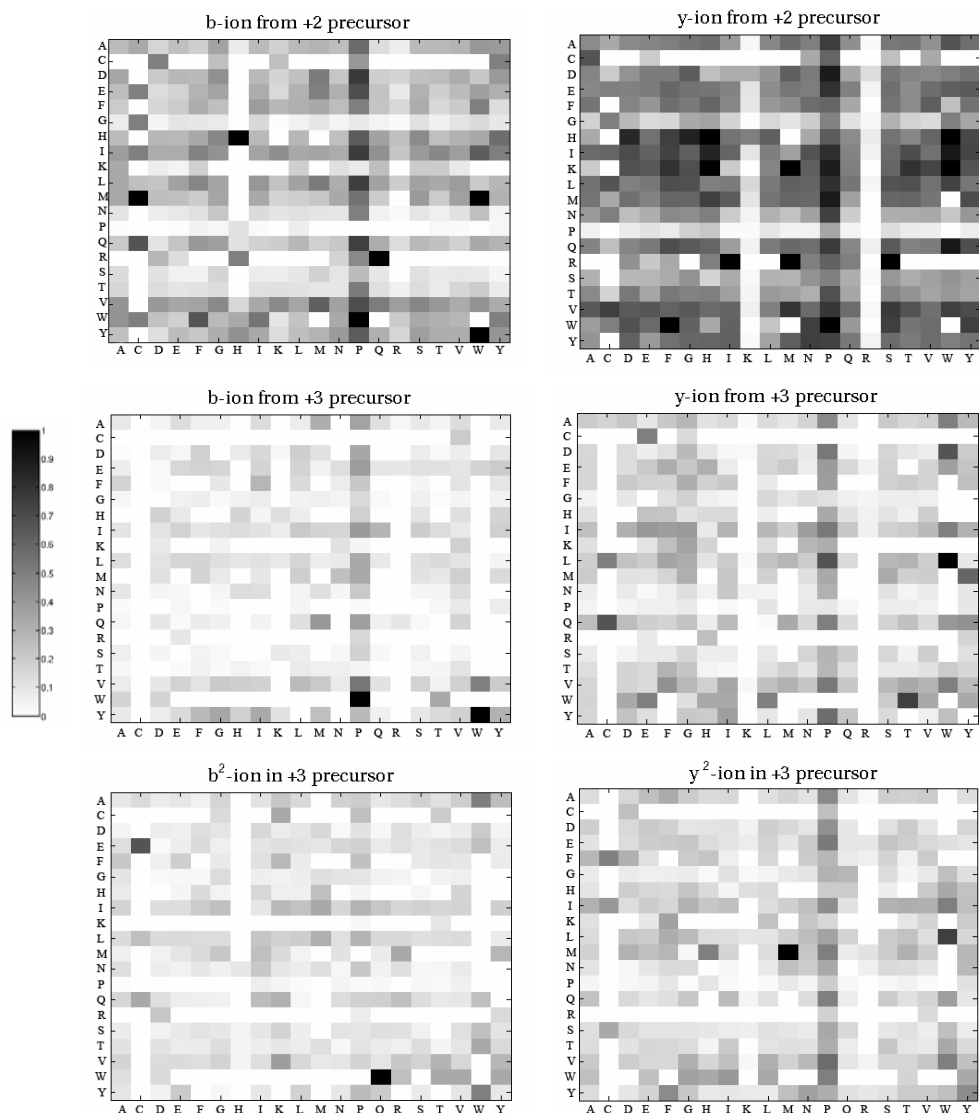
# Model selection & training

- Ensembles of two-layer feed-forward neural networks
- Experimented with network architecture
- Datasets were high-dimensional & class-imbalanced
- Different set of negatives for each network in the ensemble
- Applied feature selection & PCA
- Separate validation set for each individual model (20% of training set)
- Each ensemble contained 30 neural networks

# Performance evaluation (ROC)

<i>Ion</i>	<i>Doubly charged precursors</i>			<i>Triply charged precursors</i>		
	<i>sn</i>	<i>sp</i>	<i>acc/AUC</i>	<i>sn</i>	<i>sp</i>	<i>acc/AUC</i>
<i>precursor</i> – H <sub>2</sub> O	72.0	60.8	66.4/70.7	81.3	68.5	74.9/79.7
<i>b</i>	80.4	75.4	77.9/85.8	80.6	71.9	76.3/84.6
<i>b</i> – H <sub>2</sub> O	76.8	76.3	76.5/84.6	76.2	60.2	68.2/76.8
<i>b</i> – NH <sub>3</sub>	75.8	76.0	75.9/82.8	76.9	65.0	70.9/78.6
<i>b</i> – H <sub>2</sub> O – NH <sub>3</sub>	69.1	64.6	66.8/73.1	81.8	51.9	66.9/68.1
<i>b</i> <sup>2+</sup>	-	-	-	88.4	75.8	82.1/88.5
<i>y</i>	84.7	79.3	82.0/89.5	88.9	79.1	84.0/91.4
<i>y</i> – H <sub>2</sub> O	66.4	66.2	66.3/72.2	82.6	56.5	69.6/73.0
<i>y</i> – NH <sub>3</sub>	70.3	70.8	70.6/79.0	81.2	59.8	70.5/77.8
<i>y</i> – H <sub>2</sub> O – NH <sub>3</sub>	60.7	51.1	55.9/56.5	83.2	54.3	68.7/69.6
<i>y</i> <sup>2+</sup>	-	-	-	87.9	72.6	80.2/86.8

# Amino Acid preferences



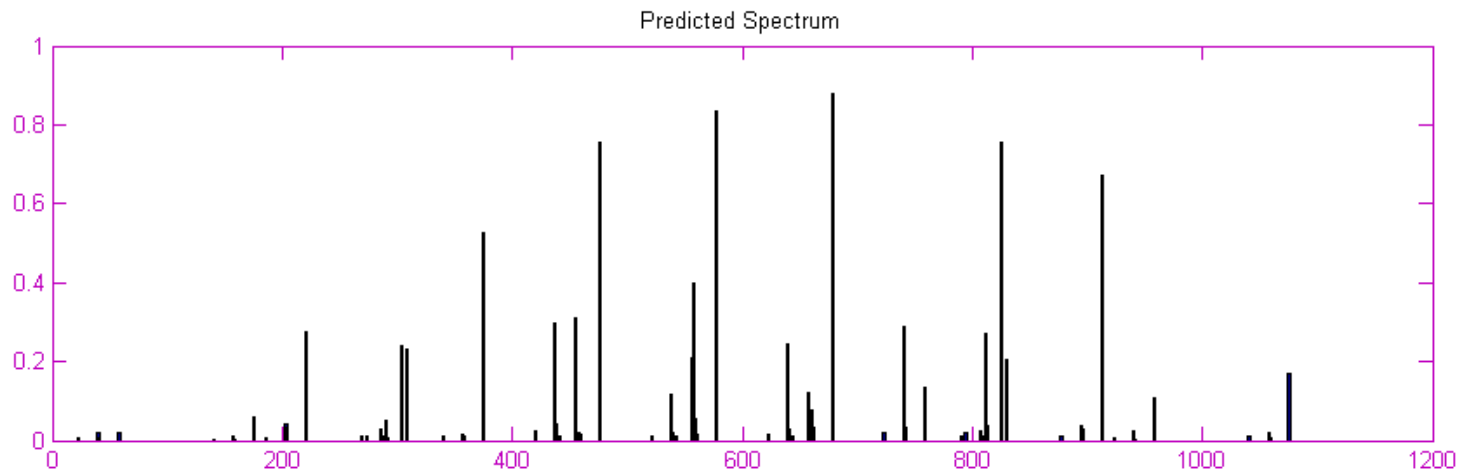
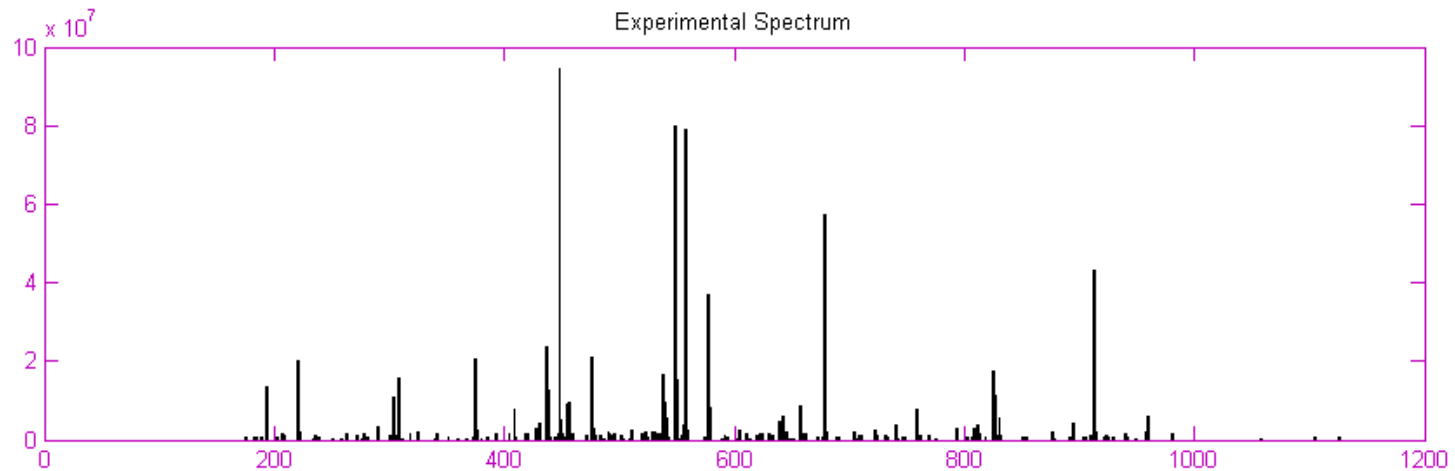
***Similar results to from studies by Smith, Wysocki, and others***

Figure 2. The amino acid preferences for peptide fragmentation. The frequencies of observing ion types (b-, y- or b<sup>2</sup>, y<sup>2</sup>) were plotted in grey scaling from 0 (white) to 1 (black). The rows indicate amino acid on the left-hand side, while the columns indicate amino acids on the right-hand side of the cleavage site.

# +2 peptides w/o Proline

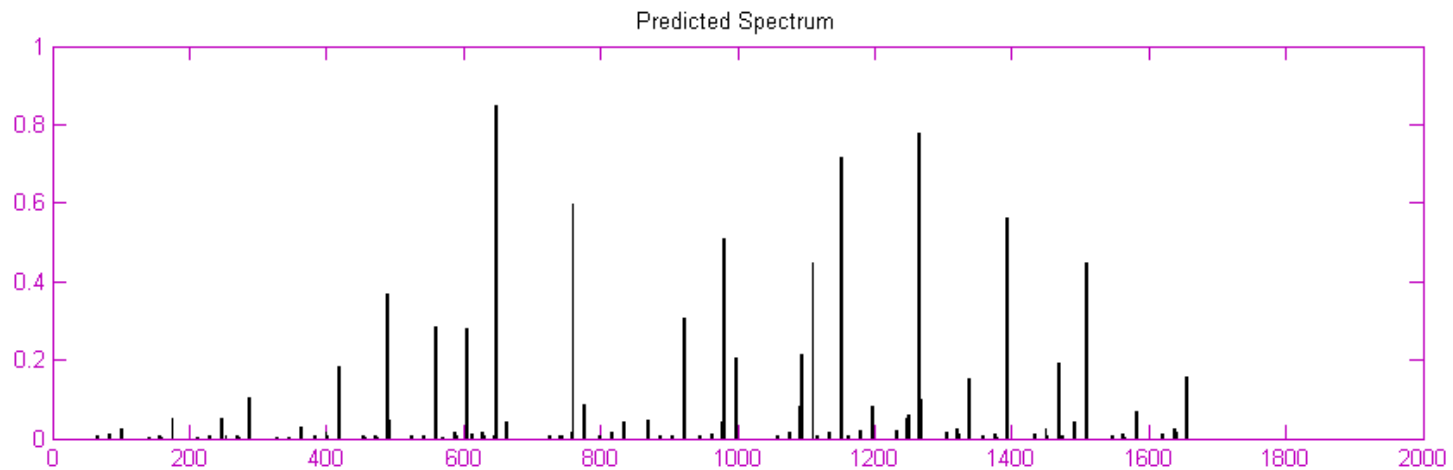
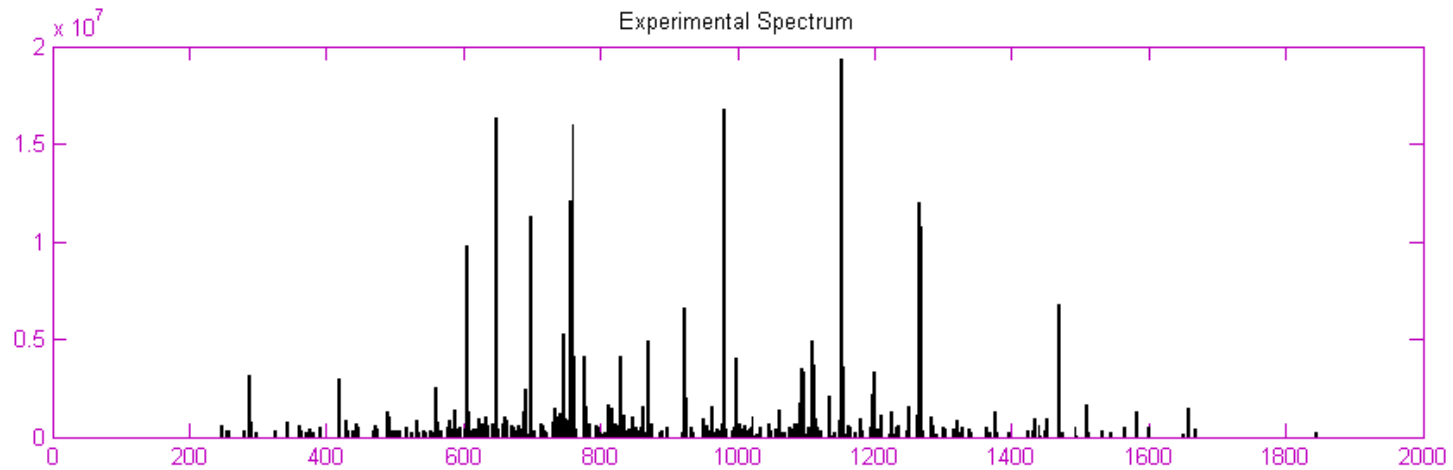
**GYSFTTTAER**

*mobile proton*



# +2 peptides w/o Proline

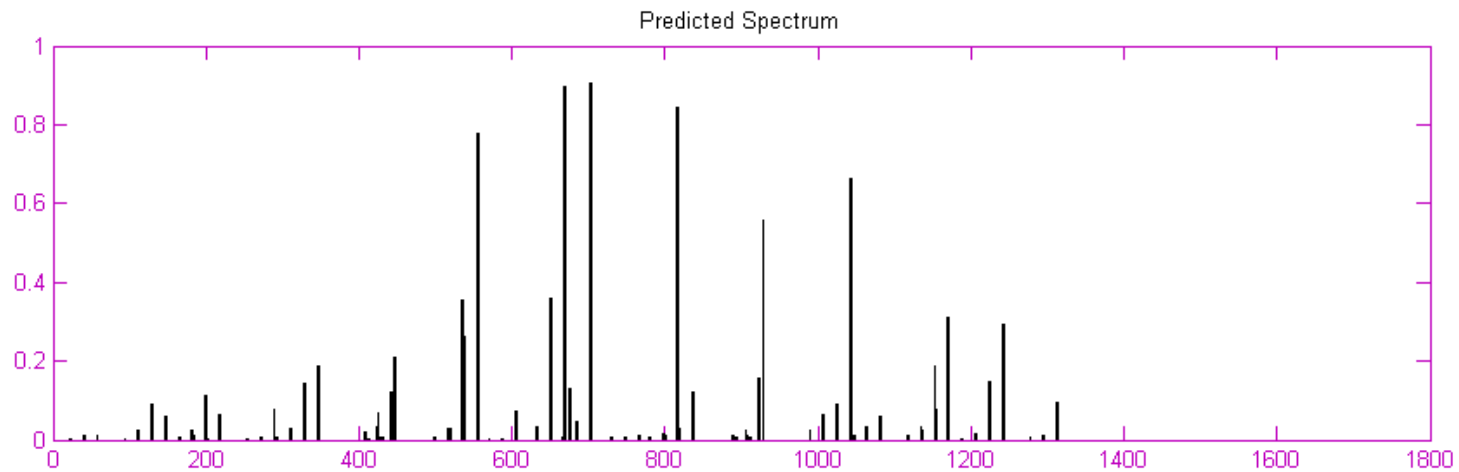
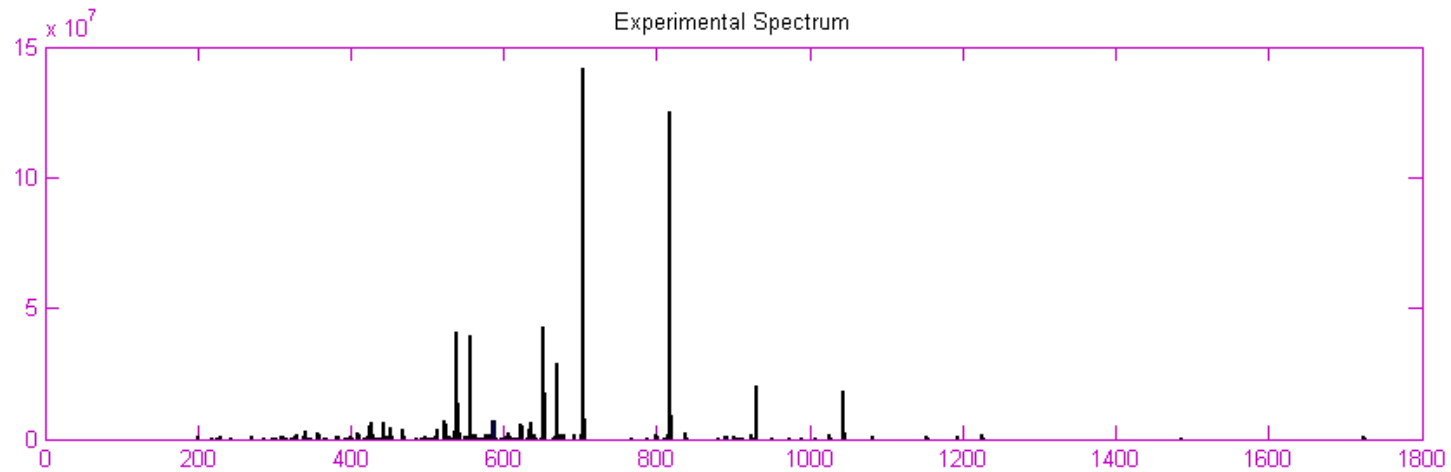
**VFDKDGNGYISAAELR**



# +2 peptides w/ Proline

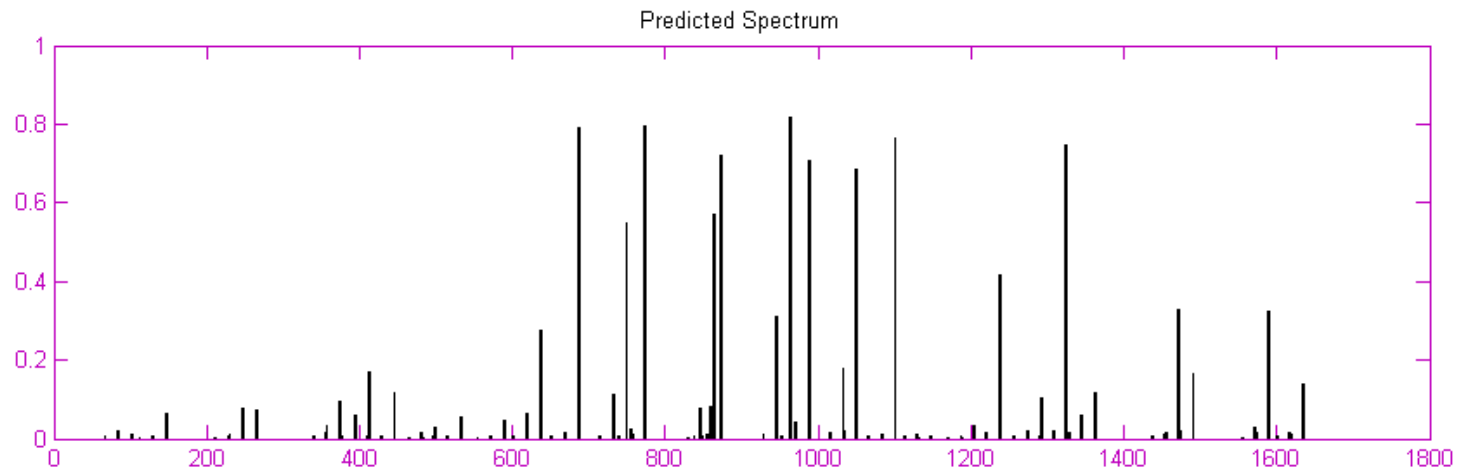
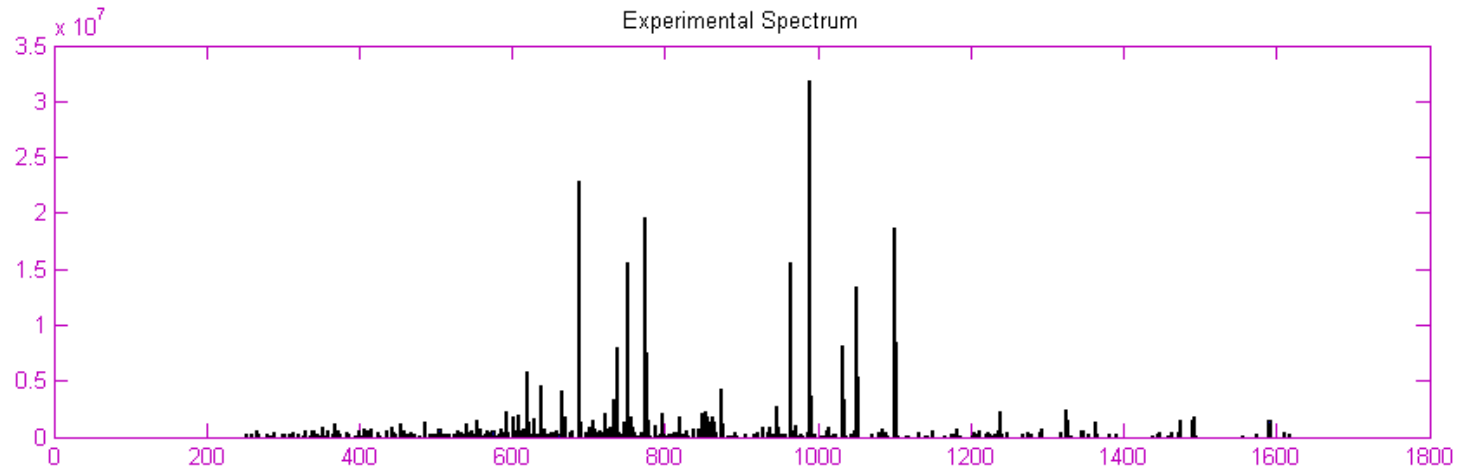
**GAAQNIIPASTGAAK**

*mobile proton*



# +2 peptides w/ Proline

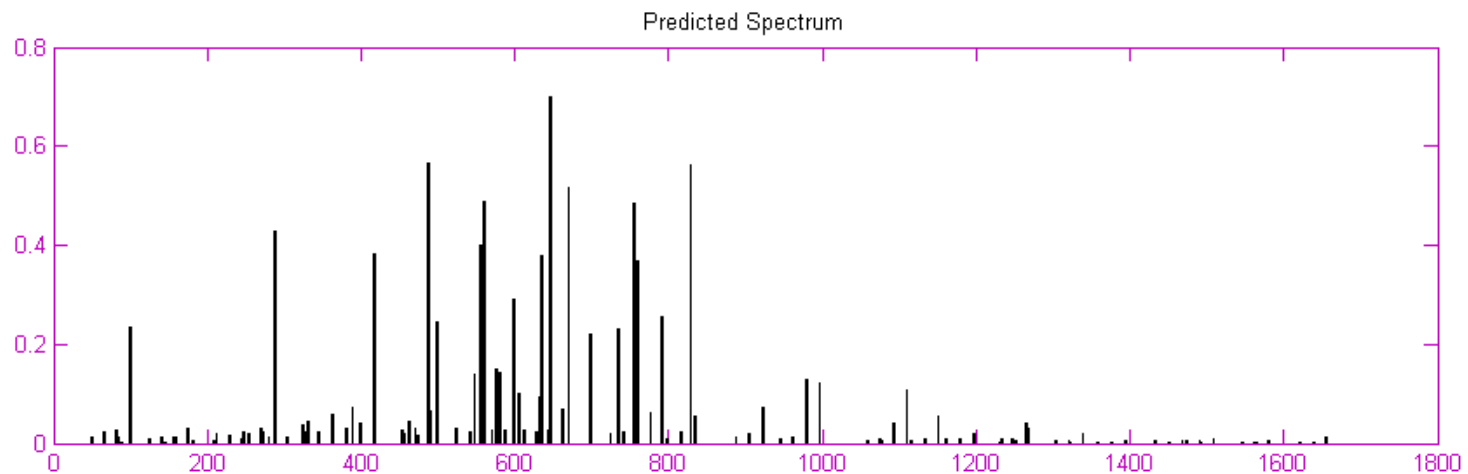
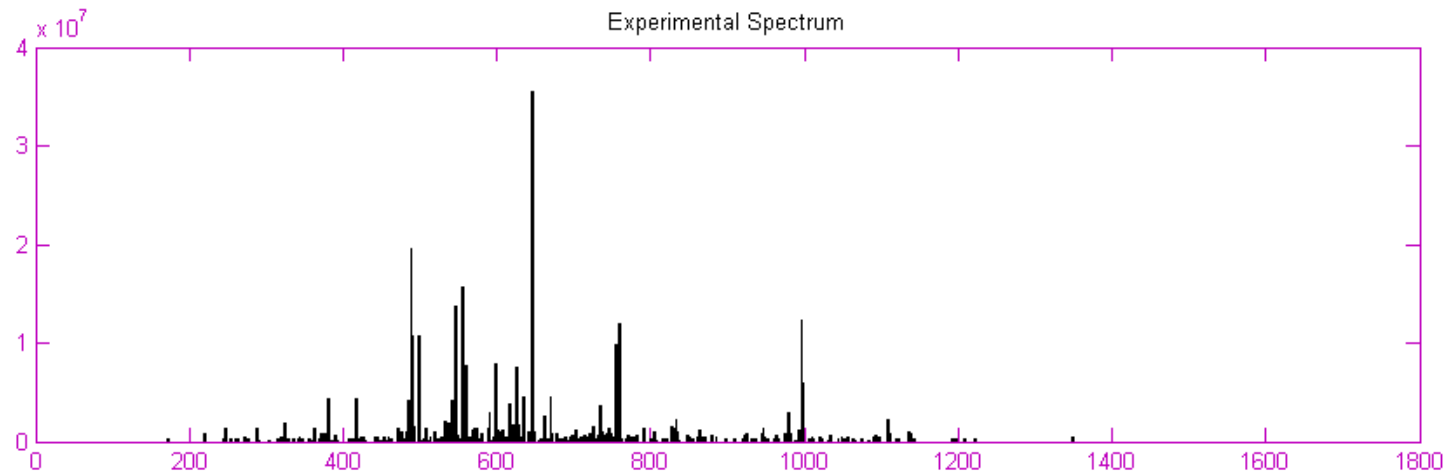
**TYFSHIDVSPGSAQVK**



# +3 peptides w/o Proline

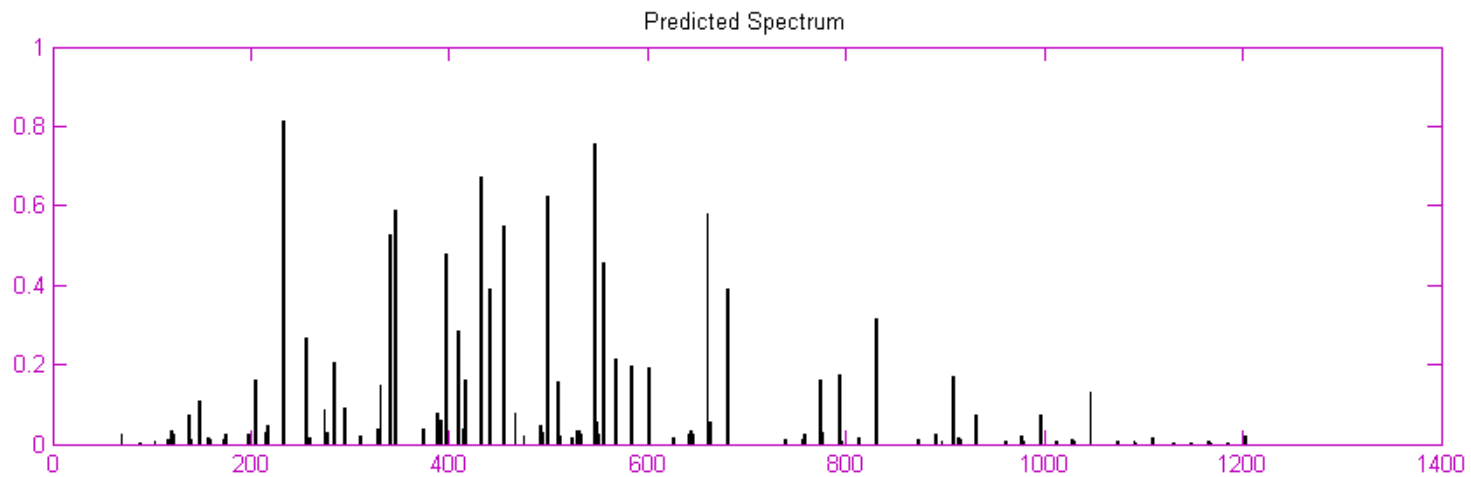
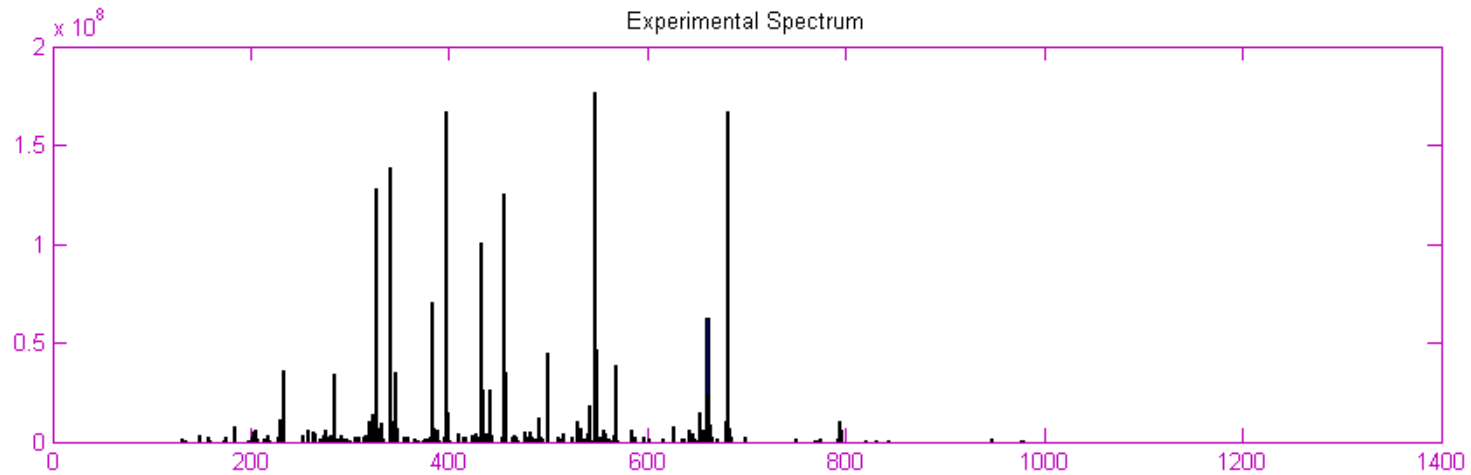
**VFDKDGNGYISAAELR**

*mobile proton*



# +3 peptides w/o Proline

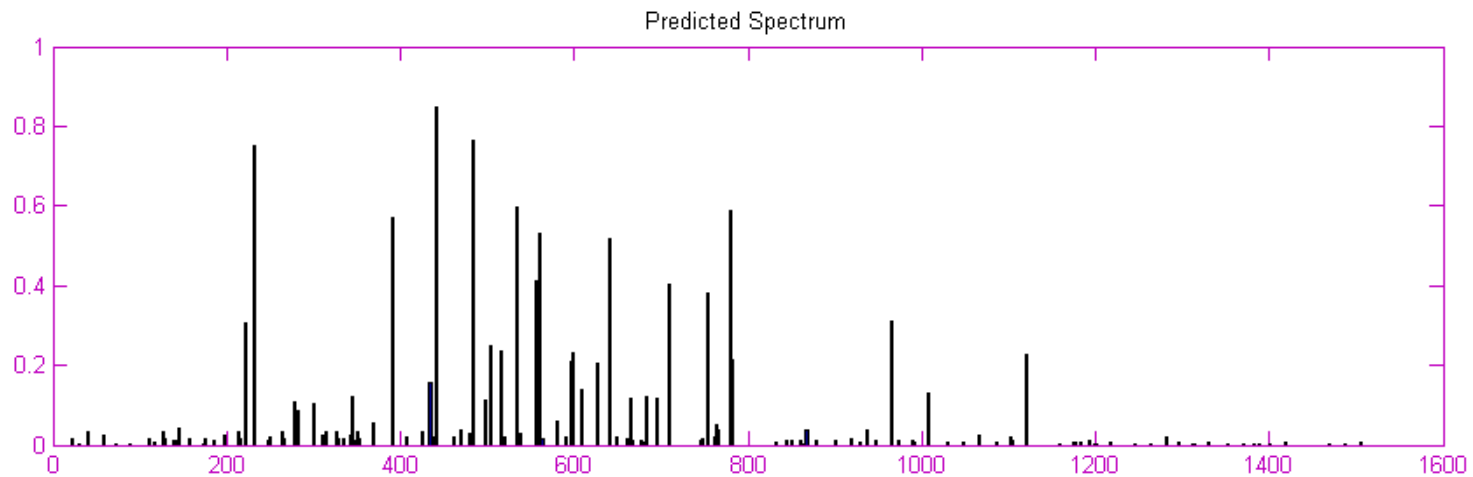
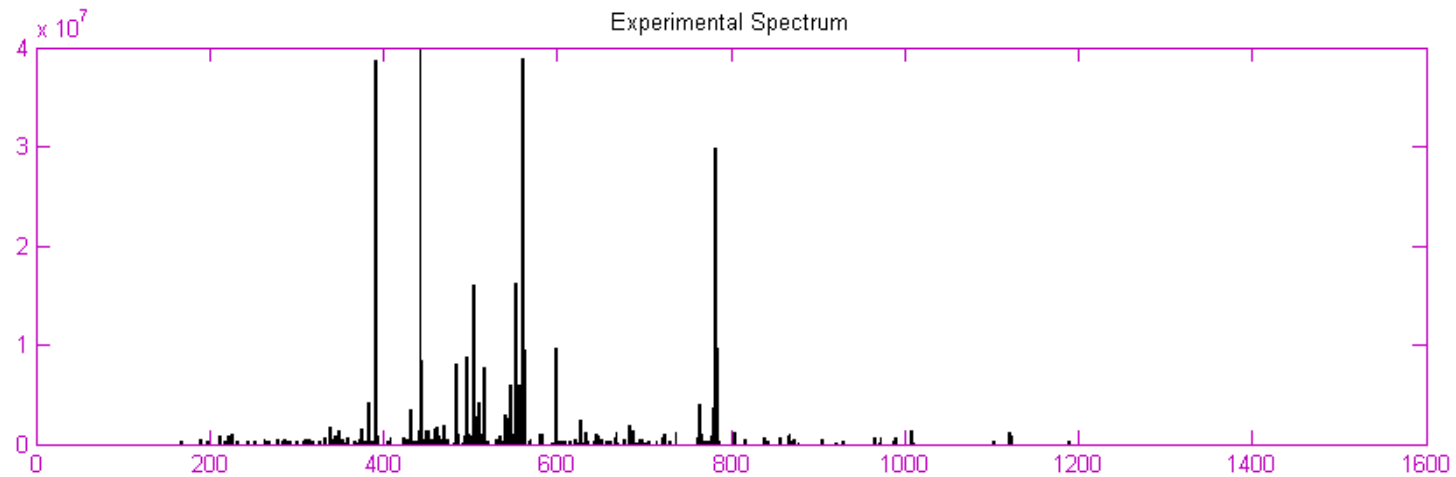
**HRDTGILDSIGR**



# +3 peptides w/ Proline

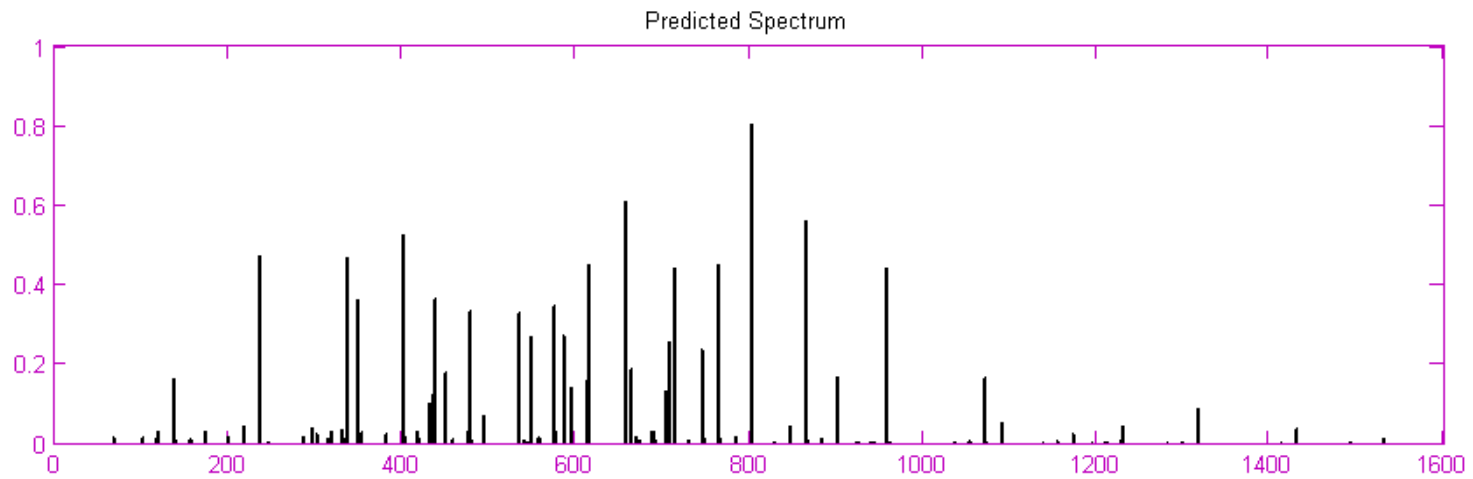
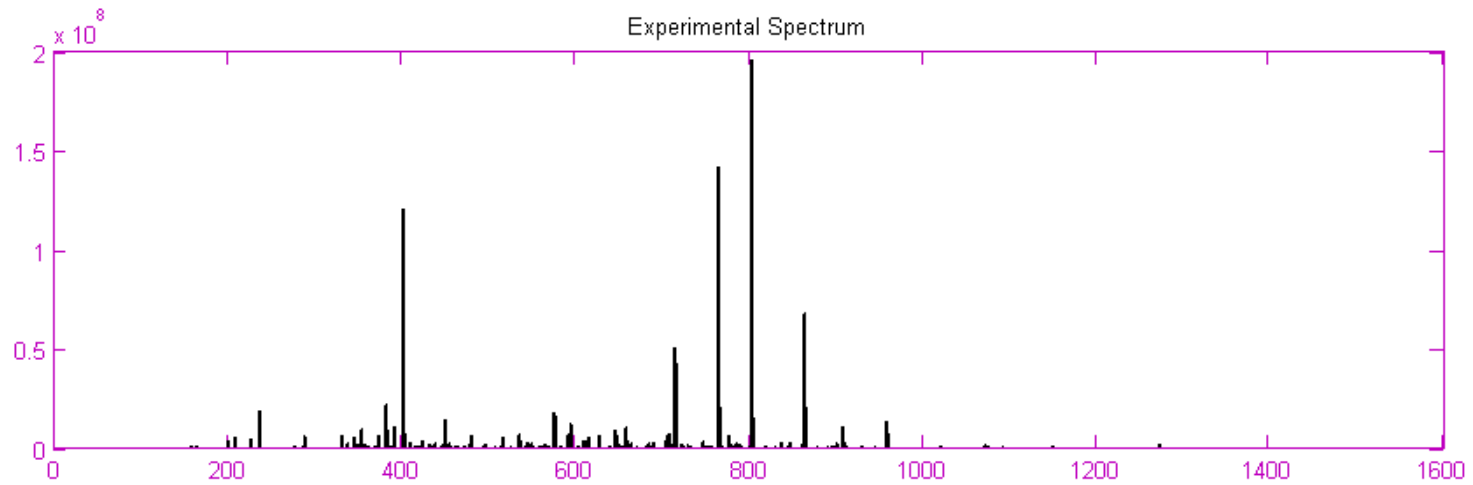
**GSHSQTSPGALPLGR**

*mobile proton*



# +3 peptides w/ Proline

**HVLSGTLGVPEHTYR**



# Peptide ID - Scoring

vs. 500 random sequences; 25 in each category

<i>Scoring scheme</i>		<i>Doubly charged precursors</i>				<i>Triply charged precursors</i>			
		<i>Mobile proton</i>		<i>Non-mobile proton</i>		<i>Mobile proton</i>		<i>Non-mobile proton</i>	
		<i>w/o Pro</i>	<i>w Pro</i>	<i>w/o Pro</i>	<i>w Pro</i>	<i>w/o Pro</i>	<i>w Pro</i>	<i>w/o Pro</i>	<i>w Pro</i>
<i>diff</i>	<i>New</i>	.32 ± .03	.26 ± .04	.30 ± .03	.24 ± .04	.13 ± .03	.14 ± .04	.22 ± .03	.25 ± .04
	<i>Simple</i>	.22 ± .02	.14 ± .03	.23 ± .02	.15 ± .02	-.01 ± .02	-.03 ± .02	.08 ± .02	.09 ± .03
<i>rank</i>	<i>New</i>	1.1 ± 0.1	1.4 ± 0.2	1.1 ± 0.1	1.5 ± 0.2	1.8 ± 0.7	1.5 ± 0.2	1.4 ± 0.4	1.2 ± 0.2
	<i>Simple</i>	1.1 ± 0.1	1.4 ± 0.2	1.0 ± 0.1	1.3 ± 0.2	9.0 ± 1.8	19.0 ± 4.5	2.3 ± 1.0	6.1 ± 2.2

# Fragmentation Prediction

## Conclusions

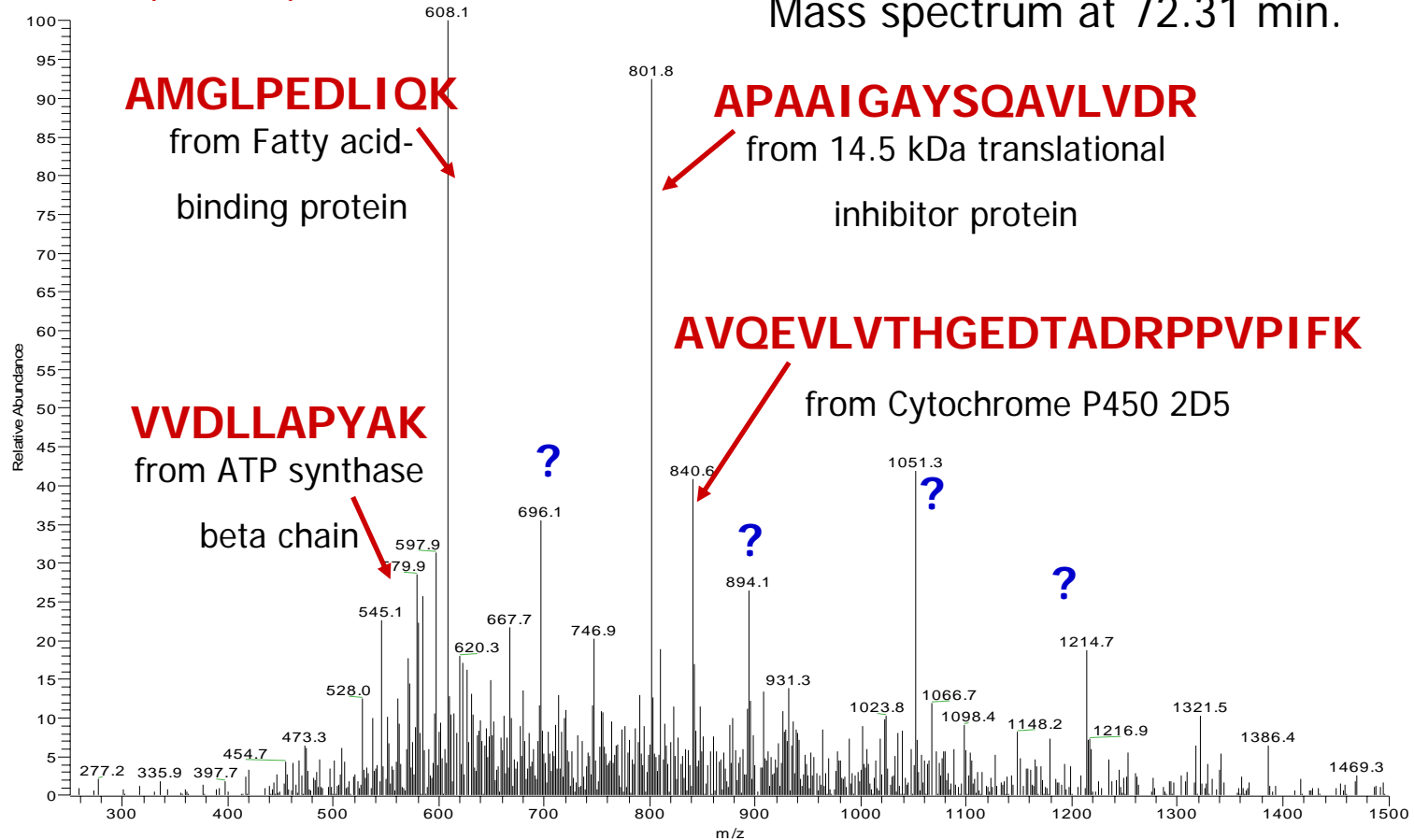
- Peptide MS/MS spectra are predictable using a neural network machine learning approach
- Observation of known mechanisms (enhanced fragmentation N-term to Pro)
- Potential to study subtle effects (mobile vs. no mobile proton)

# “Undersampling” Problem

## Incomplete proteome coverage

CH\_whole\_RG\_071503\_V06 #3291 RT: 72.31 AV: 1 NL: 9.71E8  
F: + c NSI Full ms [ 250.00-1500.00]

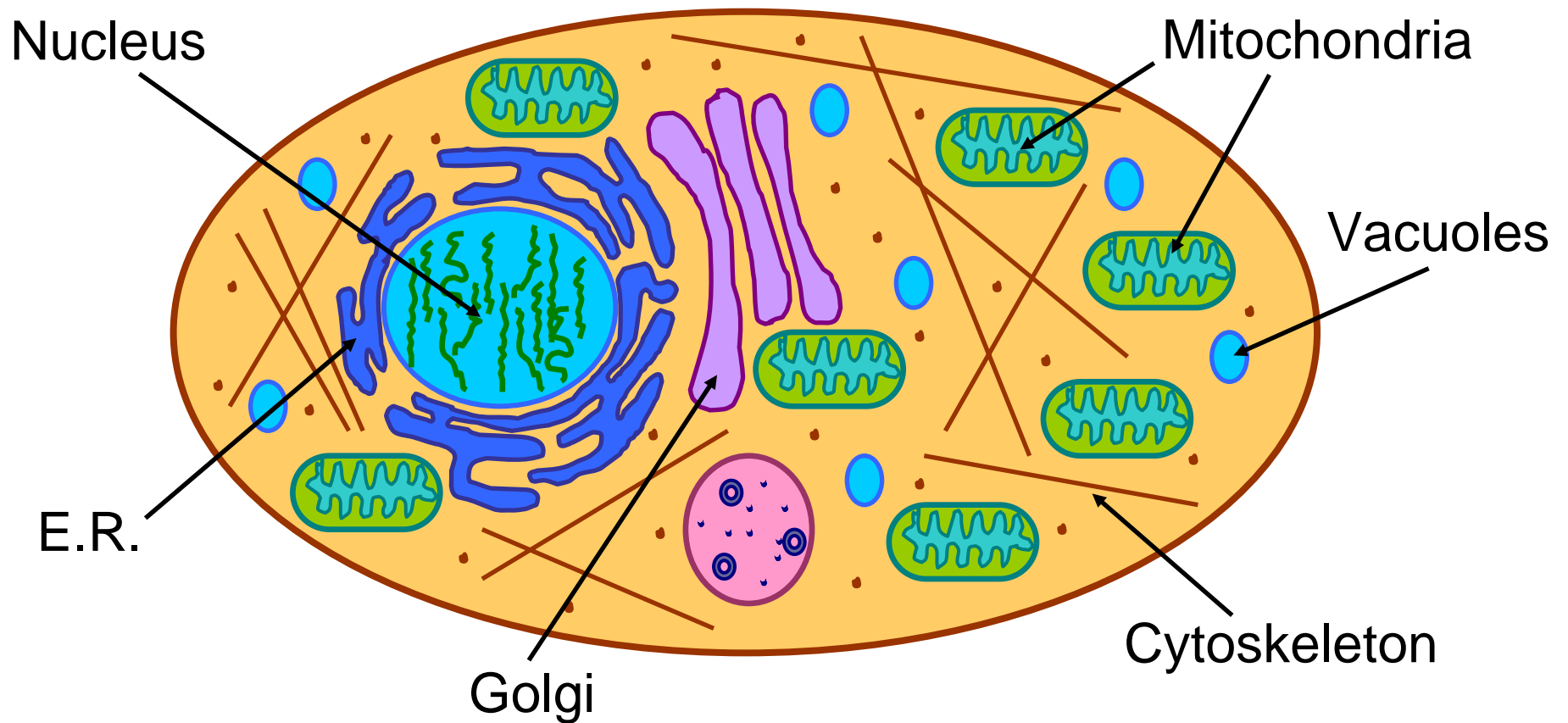
Mass spectrum at 72.31 min.



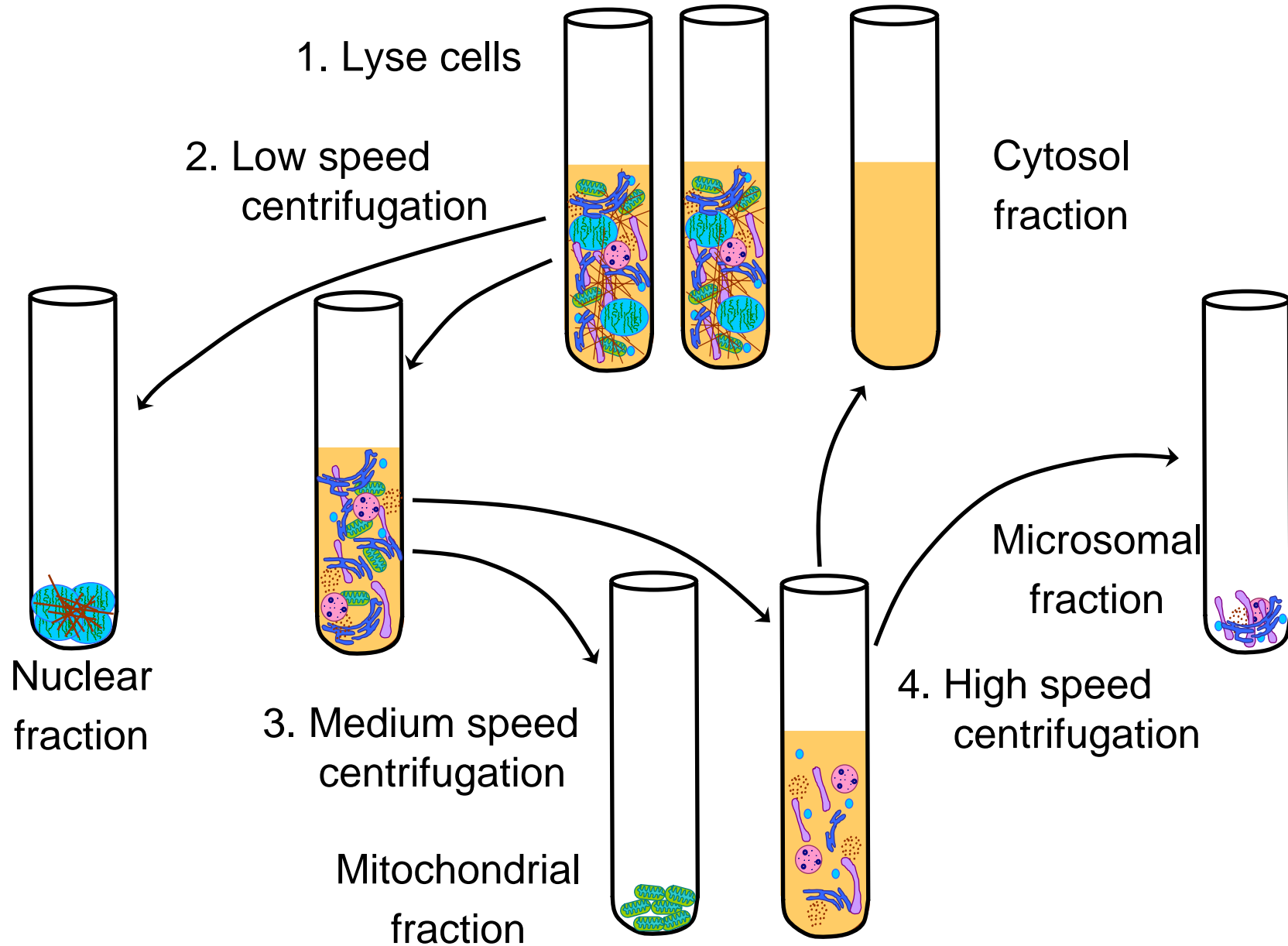
# Organelle Enrichment

Compartmentalization:

organization of eukaryotic cells into organelles



# Organelle Enrichment



# Organelle Enrichment

**Mouse Brain** tissue - Trypsin digest LC-IT-MS/MS (2x)

<u>Sample</u>	<u>Number of Peptides</u>		<u>Number of Proteins</u>	
	<u>Score &gt;28</u>	<u>1+ Peptides</u>	<u>2+ Peptides</u>	<u>3+ Peptides</u>
Whole cells	674	227	115	84
Nuclear	896	266	154	94
Mitochondrial	871	278	149	101
Microsomal	439	162	72	42
Cytosol	815	255	114	80
Combined	<b>1501</b>	<b>538</b>	<b>244</b>	<b>164</b>

# Parsing Search Results

50. Name: LDHA\_RAT                      Mass: 36427                      Score: 464                      Peptides: 10  
 Description: P04642 Lactate dehydrogenase A chain EC 1.1.1.27 LDHA LDH muscle subunit LDHM

Nuclear		Mitochondrial		Microsomal		Cytosol		peptide
<u>f1</u>	<u>f2</u>	<u>f3</u>	<u>f4</u>	<u>f5</u>	<u>f6</u>	<u>f7</u>	<u>f8</u>	
-	-	-	-	-	-	37	37	LNLVQR
35	33	45	44	-	-	57	48	LVIITAGAR
-	-	-	-	-	-	30	43	FIIPNVVK
46	43	-	-	-	-	-	-	DYSVTANSK
-	-	43	44	-	31	49	40	SADTLWGIQK
-	-	-	30	-	-	49	-	VTLTPDEEAR
-	-	-	-	-	-	43	49	QVVD SAYEVIK
-	-	-	-	-	-	62	62	AALKDQLIVNLLK + N-Ac
-	-	-	-	-	-	49	60	NVNIFKFIIPNVVK
-	-	-	-	-	-	41	-	SLNPQLGTDADKEQWK
-	-	-	-	-	-	47	-	GEMMDLQHGSFLKTPK

**Protein Results Parser 3.0**

<http://newweb.chem.indiana.edu/facilities/proteomics/parser/main.htm>

# “Cytoplasmic” Proteins

- Phosphoglycerate kinase (44.4 kDa) – glycolysis
  - Peptide Count: **0.75** / **0.5** / **0.0** / **10.25** (Nucl / Mito / Micr / Cyto)
- L-lactate dehydrogenase (36.5 kDa) – anaerobic glycolysis
  - Peptide Count: **3.0** / **2.75** / **1.5** / **11.25**
- Glyceraldehyde-3-phosphate dehyd (35.7 kDa) – glycolysis
  - Peptide Count: **4.25** / **4.0** / **10.25** / **9.0**
- Dihydropyrimidinase related protein – 2 (62.2 kDa) – axon elaboration?
  - Peptide Count: **7.5** / **8.25** / **11.75** / **16.25**
  - Subcellular Location: *tightly, but noncovalently, associated with membranes*

Data averaged from 4 LC-MS/MS analyses of two rat hippocampus tissue samples.

Protein subcellular location noted as “cytoplasmic” in SwissProt/TrEMBL (<http://us.expasy.org/>)

# Ribosomal Proteins

Peptide counts: Nucl. / Mito. / Micr. / Cyto.

- L18 (21.5 kDa) - 0.0 / 0.0 / 3.0 / 0.0
- S6 (28.7 kDa) - 0.0 / 0.0 / 2.75 / 0.0
- L12 (17.8 kDa) - 0.0 / 0.0 / 2.5 / 0.0
- 37 other proteins  
(9.3 to 47.2 kDa) 0.0 / 0.0 / 30.5 / 0.0
- SA ? (32.7 kDa) 0.0 / 0.0 / 0.0 / 2.0
- S12 (14.4 kDa) 0.0 / 0.0 / 0.0 / 0.5
- S28 (7.8 kDa) 0.0 / 0.0 / 0.0 / 0.75

# “Marker” Proteins

Peptide counts: Nucl. / Mito. / Micr. / Cyto.

- ATP synthase – mitochondrial inner membrane  
alpha (58.8 kDa) 13.75 / 13.25 / 10.0 / 0.0  
beta (56.3 kDa) 11.25 / 14.5 / 6.0 / 0.25

*incomplete separation?*

- Glutamate dehydrogenase (61.4 kDa)  
mitochondrial matrix 2.75 / 4.0 / 3.75 / 3.0

*damaged mitochondria?*

# Organelle Enrichment Conclusions

- Fast, semi-quantitation using peptide counts
- Nuclear / Mitochondrial separation challenging
- Biologically relevant information found for cytoplasmic and ribosomal proteins

# Proteomics Needs Informatics for...

- Locating peaks in 2 or more dimensions
- MS/MS spectra interpretation
- Protein/Peptide quantification
- Peptide detectability
- Experimental data → Biological information
  - enzyme or pathway regulation
  - disease susceptibility
  - drug efficacy

# Acknowledgments

## Faculty

Milos Novotny  
David Clemmer  
James Reilly  
Stephen Jacobson

## Collaborations

William McBride John Foley  
Frank Witzmann Ken Nephew  
Meei-Huey Jeng Richard DiMarchi  
Linda Malkas Martha Oakley  
Haixu Tang Pedja Radivojac  
Robert Hickey J.-T. Zhang

## Scientists / Post-docs

Yehia Mechref  
Myeong Hee Moon  
Petra Hrncirova  
Iveta Klouckova  
Steve Valentine  
Weidong Cui  
Dariusz Janecki  
Wendy Strother-Robinson  
Li-Yun Chang

## Graduate students

### Chemistry

John Taraszka  
Matt Thompson  
Arugadoss Devakumar  
Rená Sowell  
Ruwan Kurulugama  
Zhiyin (Ella) Xun

### Informatics

Kiran Annaiah  
Divya Aggarwal  
Narmada  
Jayasankar  
C.J. Fleck

### Undergraduate students

Chet Linson  
Amy Ho

## Resources

Information Technology Group

### \$\$ Funding \$\$

INGEN – Indiana Genomics Initiative  
State of Indiana 21<sup>st</sup> Century  
NIH – National Center for Glycomics  
& Glycoproteomics