Protein identification by peptide mass fingerprinting and tandem mass spectrometry

Stephen Barnes, PhD

4-7117

sbarnes@uab.edu

BMG 2-03-04

Lecture Overview

- Introduction
- Applications
- Peptide Mass Fingerprinting
  - Databases
  - homework
- Peptide fragmentation processes
  - In triple quadrupole
  - In FT-ICR cell
  - homework

Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS)

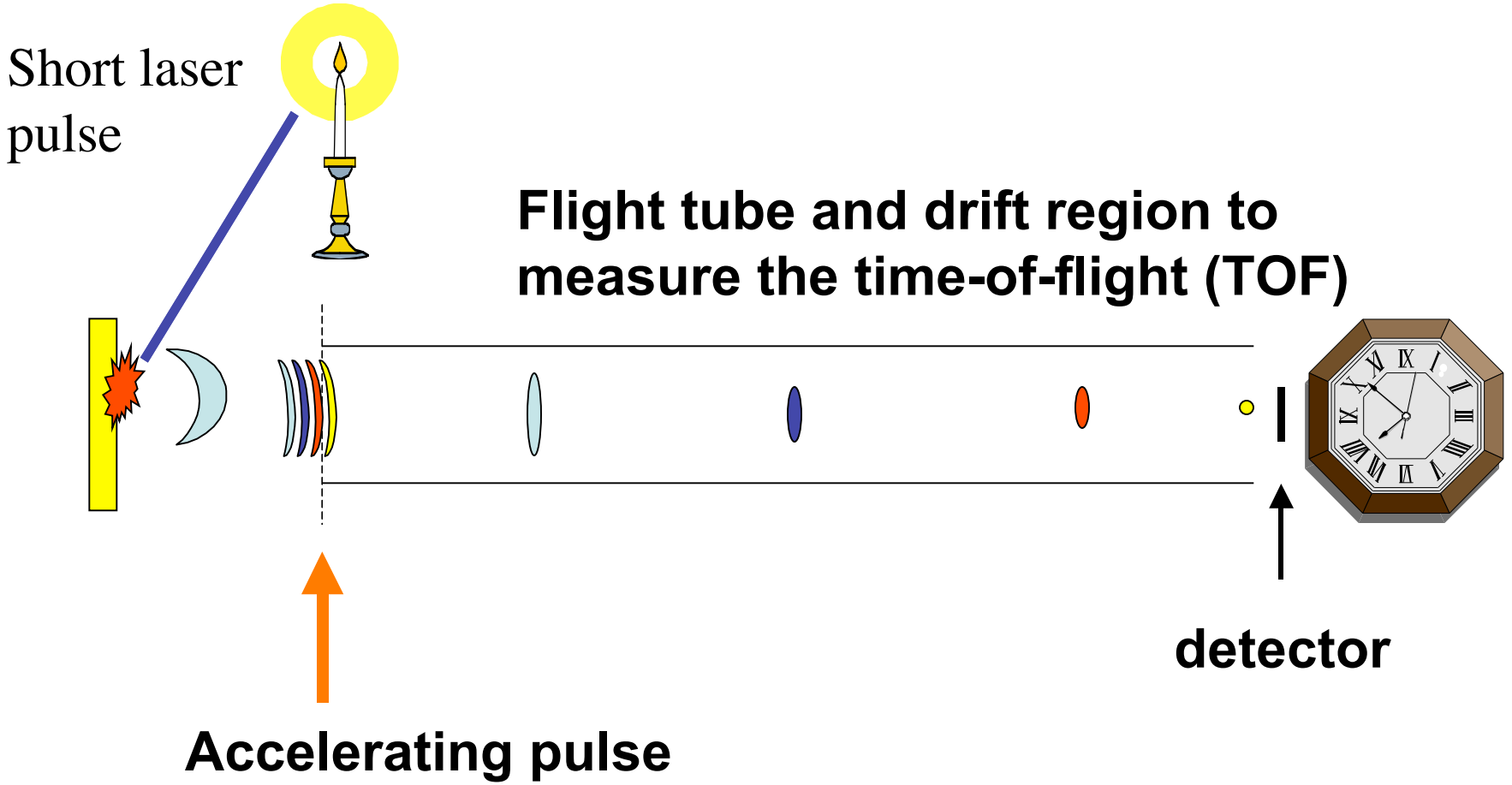
Advantages of MALDI-TOF

A) More tolerant to common buffers than ESI
B) High degree of sensitivity, moderate mass accuracy, and mass resolution
C) High mass compounds, i.e. proteins, PEG…

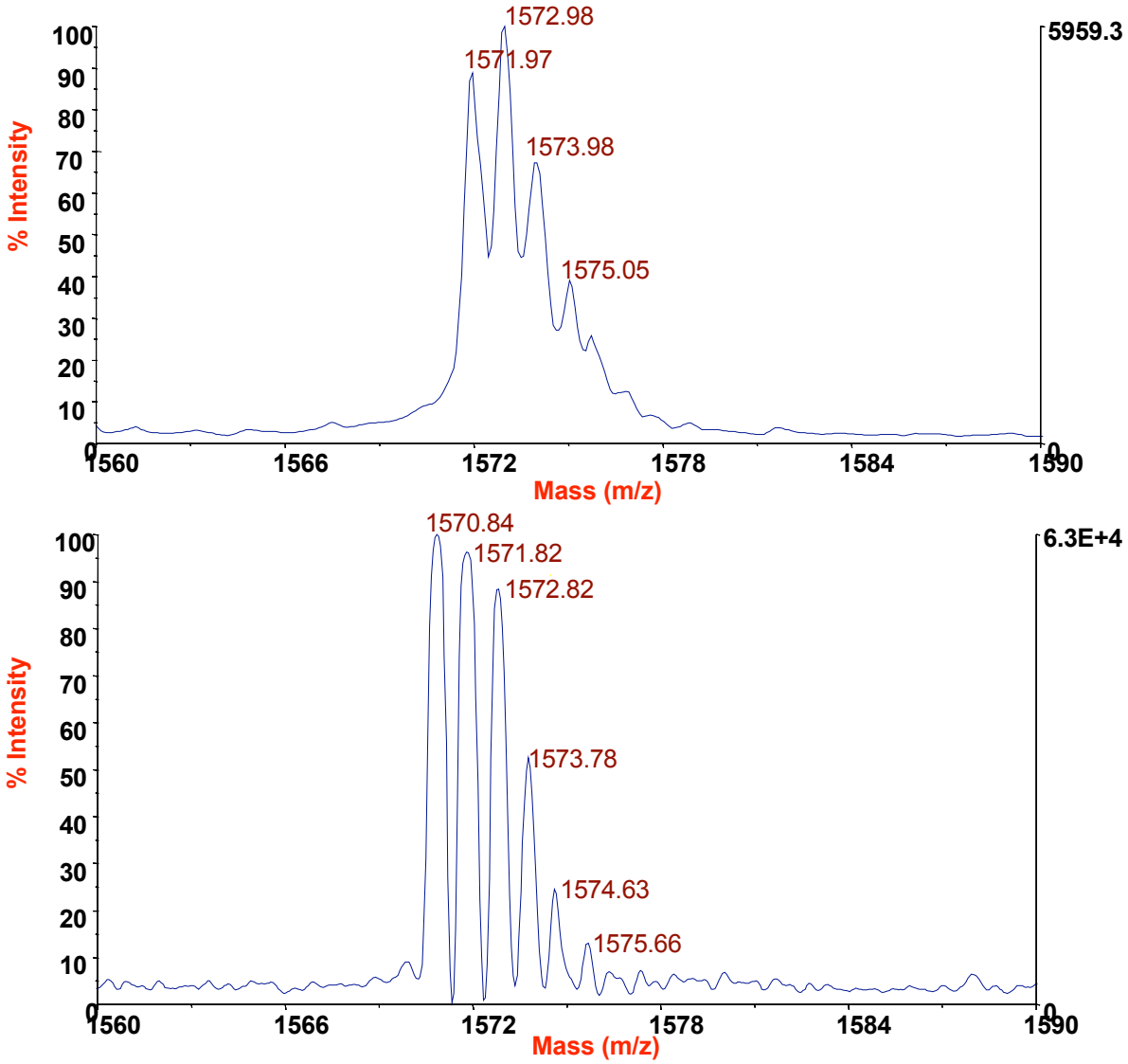
Common Applications of MALDI-TOF

A) Mass of large proteins, and other compounds
B) Enzymatic digestion profiles of proteins

Matrix-Assisted Laser Desorption Ionization (MALDI)



Increased sensitivity in reflector vs. linear mode



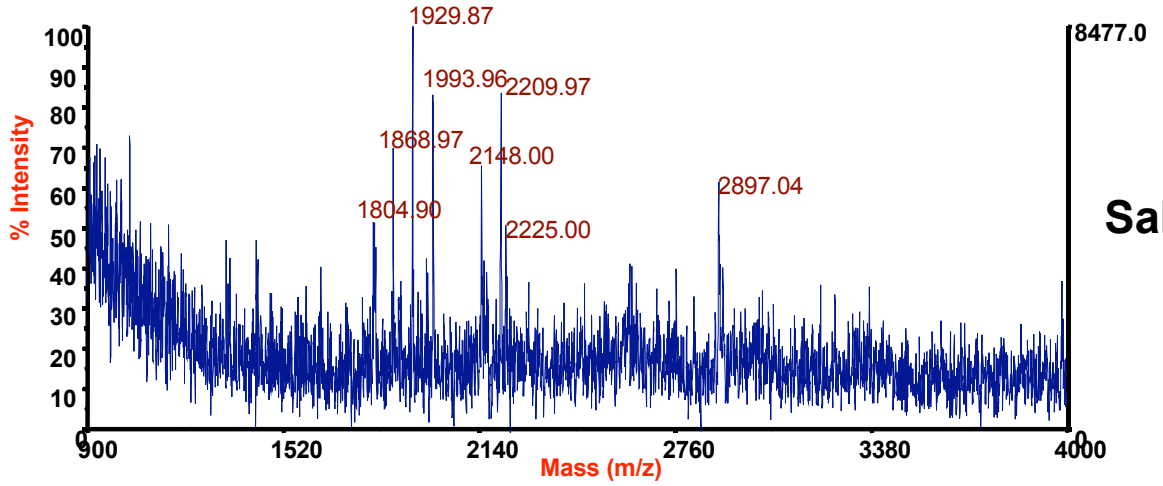
Linear Mode

Reflector Mode

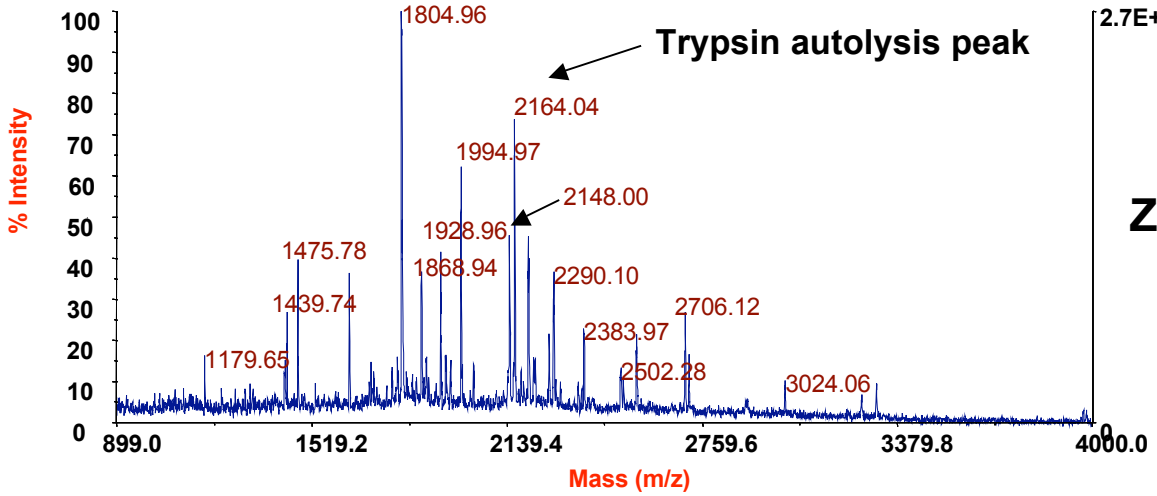
Factors from conventional experiments that impact MALDI analysis

- **Buffers used in sample preparation**
  - **NaCl up to 150 mM**
  - **Urea up to 2-3 M (carbamoylation!)**
  - **Guanidinium-HCl up to 2 M**
- **Detergents**
  - **SDS up to 0.05%**
- **Staining Protocols**
  - **Whole proteins form adducts with Coomassie**
  - **Silver staining modifies selected peptides**

Benefit of removing salt from tryptic digest



Salt Contamination



ZipTip (desalted)

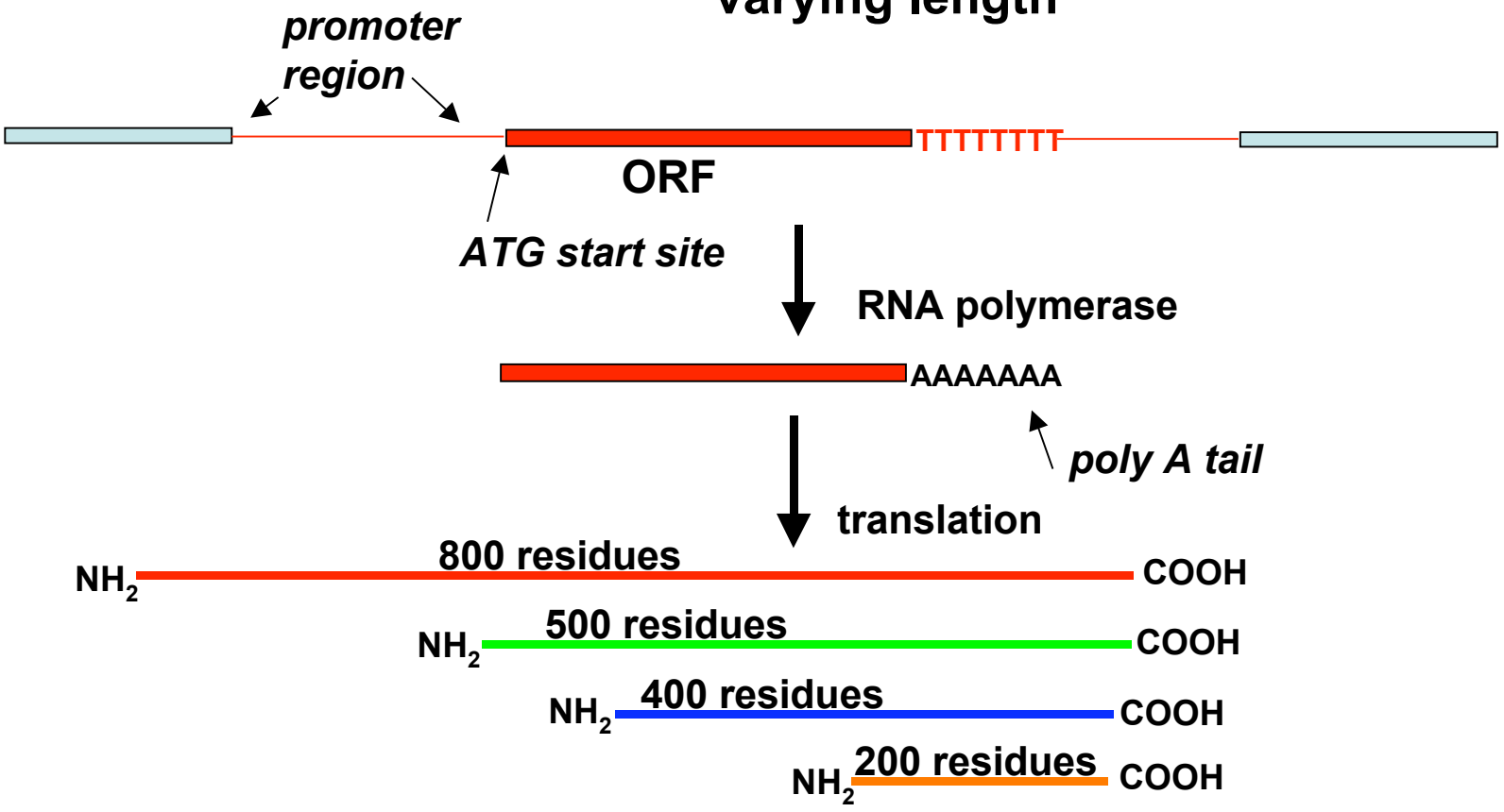
Peptide mass fingerprinting

- **This method has been developed because of the availability of predicted protein sequences from genome sequencing**
- **Proteins do not have to have been previously sequenced - only that the open reading frame in the gene is known - the rest is a virtual exercise in the hands of statisticians, bioinformaticists and computers**

From genes to proteins

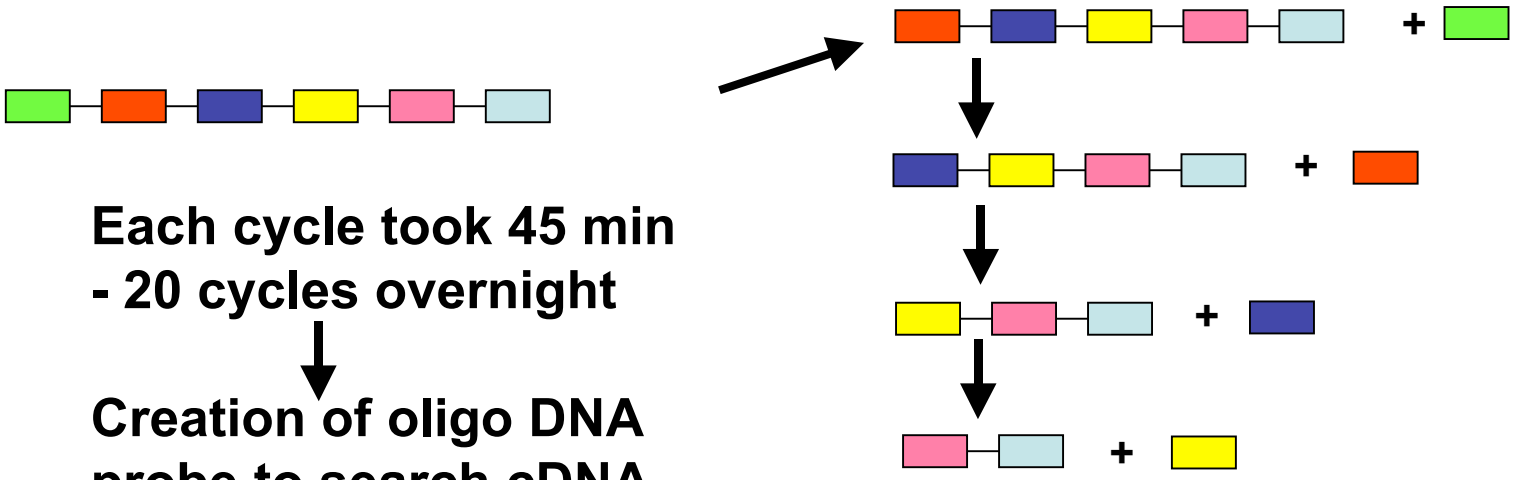
Human genome
3 billion bases

24,000+ open reading frames (ORFs), each encoding an individual protein sequence of varying length



Sequencing proteins
pre-1995

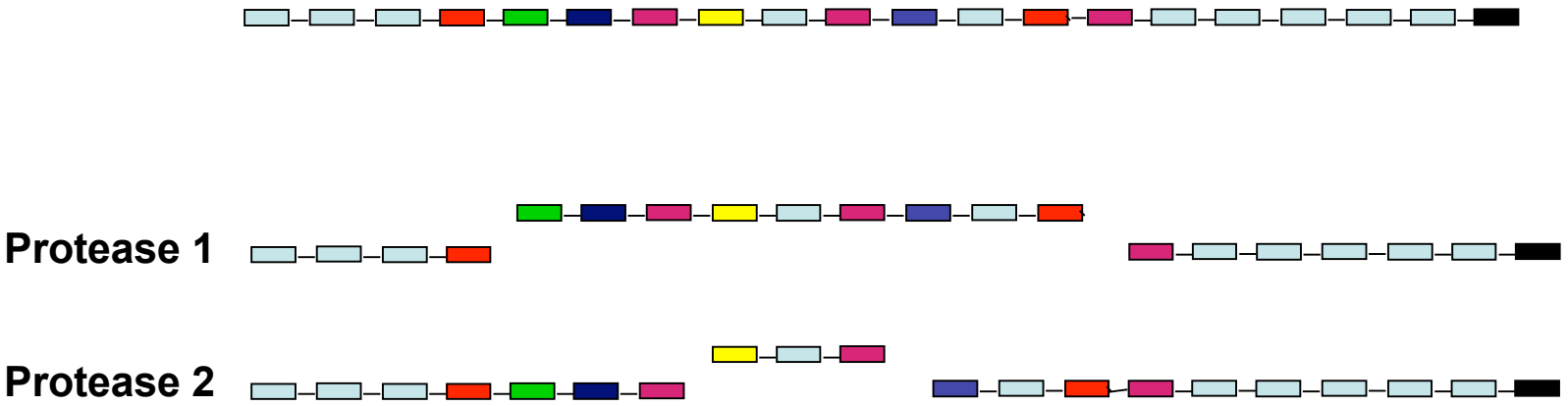
N-terminal amino acid sequencing of an isolated polypeptide by a chemical procedure *(Edman degradation)* without fragmentation of the polypeptide



Each cycle took 45 min
- 20 cycles overnight

Creation of oligo DNA probe to search cDNA library

Protein sequencing by overlapping peptide sequences



Sequencing of DNA was carried out using a similar strategy

From DNA to peptide fragments

.ATG.CTT.CCT.CAC.GGT.AAA.TCG.TAT.GCT….

NH_2-Met.Leu.Pro.His.Gly.Lys.Ser.Tyr.Ala….

Trypsin

NH_2-Met.Leu.Pro.His.Gly.Lys-COOH

From Proteins to Sequence Tags

- If each protein (average 500 residues) had a cleavage site every 10 residues, then about 1.5 million peptides describe the expressed products of the human genome
- Each peptide has a molecular weight value that is its individual sequence tag
- Any modification will increase the peptide's molecular weight

Peptide information needed for protein identification

- **Peptide-mass fingerprinting and the ideal covering set for protein characterization. M. Wise et al. *Electrophoresis* 18:1399-1409, 1997**

- **Purpose: To determine the efficiency and nature of protein identification by the use of endoproteinases and mass spectrometry to create and identify the resulting peptides**

Setup

Database of 128,719 non-redundant protein entries

Assumptions:

1. **Digestion is always perfect (value of being *in silico*)**
2. **Cleavage always occurs on the carboxy terminal of each amino acid**
3. **Fragment masses were accurate to the nearest dalton, i.e., ± 0.5 Da**

Theoretical proteolysis of derived protein database

In silico **endoproteinases**

- **All possible single amino acid sites**
- **Biochemical endoproteinases: chymotrypsin, trypsin and Glu-C**

Results for chymotrypsin

Database entries:	**128,719**
# of peptide fragments:	**3,086,608**
# of distinct fragments	**14,778**
Size of largest fragment:	**243,718 Da**
Max # of entries for a particular fragment	**20,926 (260 Da)**
Average # entries for a given fragment:	**209**
Average number of fragments for an entry:	**24**
# of uncut entries:	**3,059**
Average size of uncut entries:	**3,194 Da**
Max size of uncut entry:	**65,243 Da**

# of entries defined by X fragments

X=1:	**2,900**
X=2:	**88,118**
X=3:	**26,369**
X=4:	**952**
X=5:	**48**
X=6:	**13**
X=7:	**2**
X=8:	**1**
X=9:	**1**

Average # of fragments to define a protein: 2.216

Summary of digestion data

Amino acid	Distinct Fragments	Avg # fragments	#Uncut	Avg ident
A alanine	15,372	21.45	3,468	2.13
C cysteine	38,661	6.40	21,525	1.91
D aspartate	17,163	16.15	6,936	2.05
E glutamate	16,960	18.43	6,555	2.08
F phenylalanine	21,642	12.92	7,788	2.00
G glycine	16,490	20.42	3,531	2.13
H histidine	28,695	7.72	18,104	1.96
I isoleucine	18,227	17.36	6,735	2.08
K lysine	19,821	17.50	6,673	2.07
L leucine	12,490	26.19	3,598	2.23
M methionine	29,873	7.88	14,409	1.95
N asparagine	19,765	14.41	8,077	2.03
P proline	19,437	15.34	6,590	2.04
N glutamine	20,182	12.84	8,062	2.01
R arginine	18,754	16.07	6,633	2.07
S serine	13,829	21.51	3,446	2.15
T threonine	15,455	18.21	4,451	2.11
V valine	15,089	19.61	5,084	2.11
W tryptophan	39,643	5.09	26,214	1.91
Y tyrosine	24,343	10.79	9,738	1.98
Glu-C	11,291	30.88	2,808	2.28
Chymotrypsin	14,780	25.42	2,822	2.22
Trypsin	10,846	30.37	2,418	2.34

Protein analysis by MALDI 2004

Eppendorf tube

destain

Speed-Vac

trypsin

1:20

Incubate

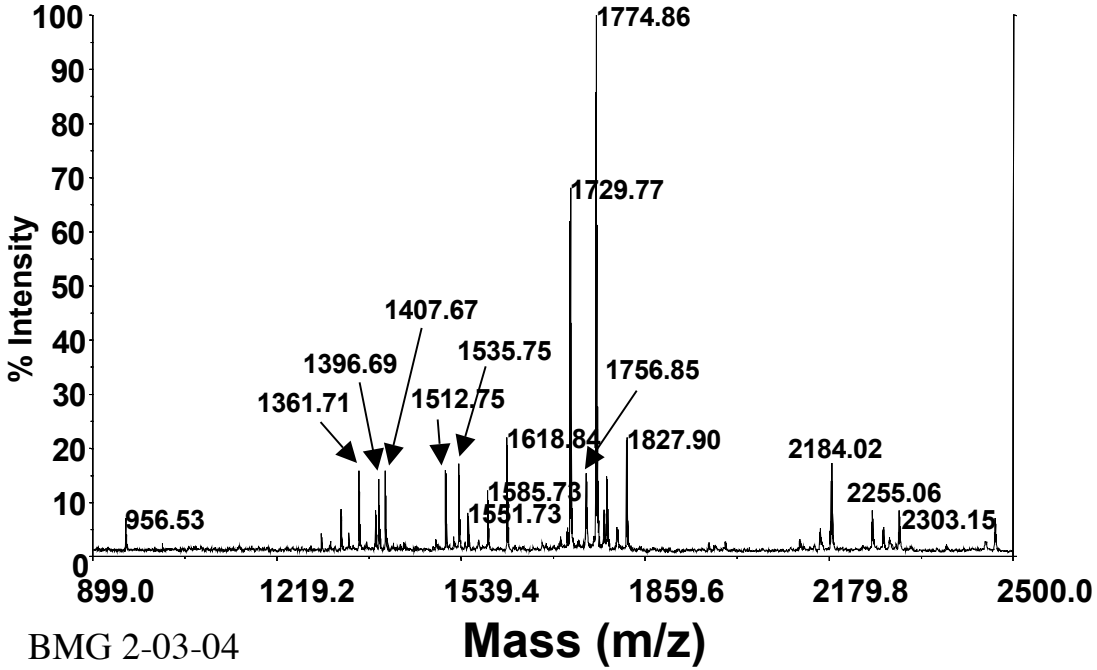
overnight

Water Bath 37°C

Peptide extraction

MALDI plate

Desalting Ziptip



Choice of peptidase

- Analogous to DNA restriction enzymes
- Tryptic peptide fingerprinting may identify several related protein candidates (e.g., actins)
- Inspection of the sequences may reveal that there is a difference at one residue that distinguishes between two candidates.
- If for instance it is a glutamate, then use of Glu-C or V8-protease may enable the two proteins to be correctly identified
- INSPECT sequences carefully

Proteolytic enzymes used to hydrolyze proteins

The choice of enzyme largely depends on the nature of the amino acid sequence and the specific issue that is being addressed

- **Trypsin - *cleaves at arginine and lysine residues***
- **Chymotrypsin - *cleaves hydrophobic residues***
- **Arg-C - *cleaves at arginine residues***
- **Glu-C - *cleaves at glutamic acid residues***
- **Lys-C - *cleaves at lysine residues***
- **V8-protease - *cleaves at glutamic acid residues***
- **Pepsin - *cleaves randomly, but at acid pH***

***See* http://www.abrf.org/JBT/1998/September98/sep98m_r.html**

Genomics and proteins in 2004

- **The human genome consists of about 30,000 genes that are expressed as proteins**
- **Large Scale Biology Corp has cataloged 116,000+ protein forms from human tissues, representing the expressed products of 18,000 genes**
- **The expected number of protein forms is expected to be in excess of 200,000**

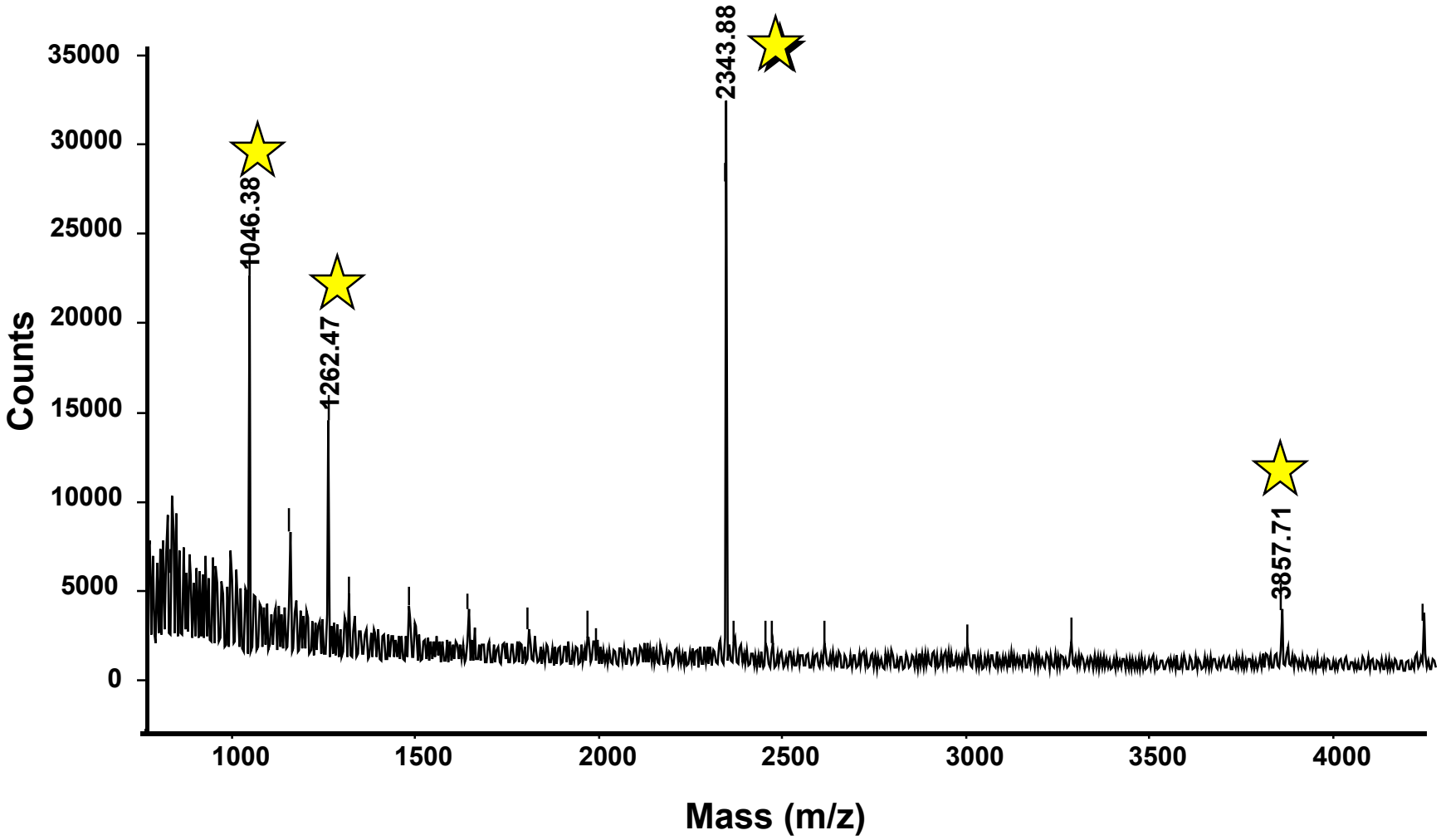
Searching databases with peptide masses to identify proteins

Best site is at www.matrixscience.com

The program (MASCOT) can search the OWL or NCBI databases using a set of tryptic peptide masses, or the fragment ions (specified or unspecified) of peptides

Presents the expected set of tryptic peptides for each matched protein

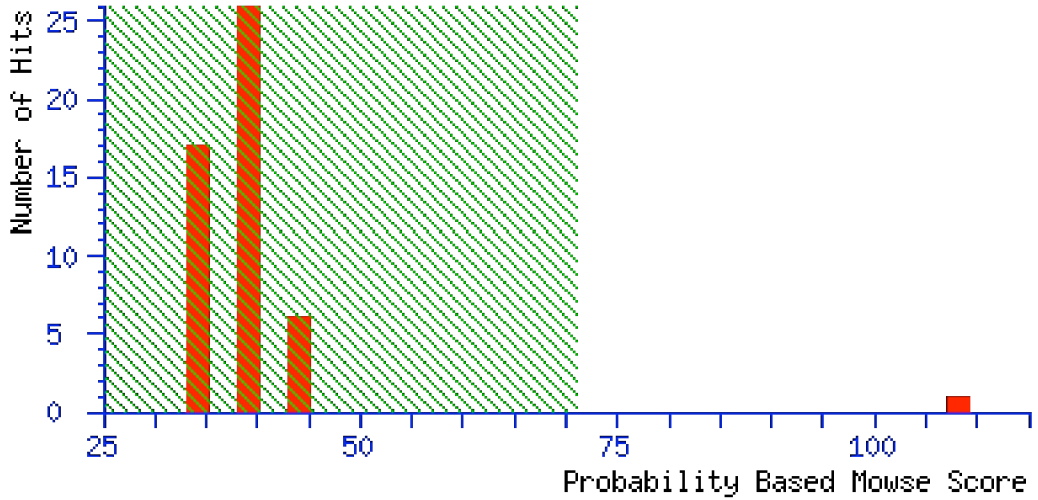
MALDI-TOF mass spectrum of tryptic digest of p22 band purified by 6xHis-tag



Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event.

Protein scores greater than 71 are significant (p<0.05).



Accession	Mass	Score	Description
1. gi|548939	20840	108	FKBP-TYPE PEPTIDYL-PROLYL CIS-TRANS ISOMERASE SLYD (PPIASE) (ROTAMA
2. gi|13384624	46931	45	myocyte enhancer factor 2C [Mus musculus]
3. gi|5257384	43424	44	(AF137308) phytochrome B [Lolium perenne]
4. gi|4505147	50305	44	MADS box transcription enhancer factor 2, polypeptide C (myocyte enhan
5. gi|1515365	44552	43	(U52596) nucleocapsid protein [Avian infectious bronchitis virus]
6. gi|6093850	49443	42	PRESENILIN 2 (PS-2)
7. gi|15225198	47999	42	hypothetical protein [Arabidopsis thaliana]
8. gi|113854	58376	41	NITROGENASE IRON-IRON PROTEIN ALPHA CHAIN (NITROGENASE COMPONENT I)
9. gi|13928425	13831	40	(AB040419) envelope protein [Bovine immunodeficiency virus]
10. gi|4389228	56064	40	Chain Z, Crystal Structure Of The Complex Between Escherichia Coli Glycerol

MASCOT SEARCH SUMMARY

1. gi|548939 Mass: 20840 Score: 108
FKBP-TYPE PEPTIDYL-PROLYL CIS-TRANS ISOMERASE SLYD (PPIASE) (ROTAMA

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
1046.38	1045.37	1045.59	-0.22	132 -	140	0	FNVEVVAIR
1262.47	1261.46	1261.70	-0.24	6 -	16	0	DLVVSLAYQVR
2343.88	2342.87	2343.08	-0.20	58 -	78	0	FDVAVGANDAYGQYDENLVQR
3857.71	3856.70	3856.89	-0.19	96 -	131	0	FLAETDQGPVPVEITAVEDDHVVVDGNHMLAGQNLK

2. gi|13384624 Mass: 46931 Score: 45
myocyte enhancer factor 2C [Mus musculus]

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
1046.38	1045.37	1045.50	-0.13	263 -	271	0	NTMPSVNQR
3857.71	3856.70	3856.76	-0.06	178 -	218	0	NSMSPGVTHRPPSAGNTGGLMGGDLTSGAGTSAGNGYGNPR

No match to: 1262.47, 2343.88

3. gi|5257384 Mass: 43424 Score: 44
(AF137308) phytochrome B [Lolium perenne]

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
1046.38	1045.37	1045.54	-0.17	380 -	389	0	GIDELSSVAR
3857.71	3856.70	3856.72	-0.02	86 -	122	0	SPHGCHAQYMANMGSIASLVMAVIISSGGEDEHNMGR

No match to: 1262.47, 2343.88

4. gi|4505147 Mass: 50305 Score: 44
MADS box transcription enhancer factor 2, polypeptide C (myocyte enhan

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
1046.38	1045.37	1045.50	-0.13	265 -	273	0	NTMPSVNQR
3857.71	3856.70	3856.76	-0.06	180 -	220	0	NSMSPGVTHRPPSAGNTGGLMGGDLTSGAGTSAGNGYGNPR

No match to: 1262.47, 2343.88

E. coli: FKBP-TYPE PEPTIDYL-PROLYL CIS-TRANS ISOMERASE

Nominal mass of protein (Mr): 20840

```
  1 MKVAKDLVVS LAYQVRTEDG VLVDESPVSA PLDYLHGHGS
 41 LISGLETALE GHEVGDKFDV AVGANDAYGQ YDENLVQRVP
 81 KDVFMGVDEL QVGMRFLAET DQGPVPVEIT AVEDDHVVVD
121 GNHMLAGQNL KFNVEVVAIR EATEEELAHG HVHGAHDHHH
161 DHDHDGCCGG HGHDHGHEHG GEGCCGGKGN GGCGCH
```

Tryptic fragments detected by MALDI-TOF-MS

```
132-140 FNVEVVAIR
  6- 16 DLVVSLAYQVR
 58- 78 FDVAVGANDAYGQYDENLVQR
 96-131 FLAETDQGPVPVEITAVEDDHVVVDGNHMLAGQNLK
```

Other web sites for peptide analysis

- **http://prowl.rockefeller.edu/**
  - **Choose ProFound**
- **http://prospector.ucsf.edu/**
  - **Choose MS-fit**

Further information on identified protein

- **Take the protein identifier number:**
  - **For this example it is gi|548939**
  - **Go to http://www.ncbi.nlm.nih.gov**
  - **Under Entrez, paste in the gi number**
  - **A link to the protein will appear**
  - **Click on Blink - this is similar to BLAST, but better**
  - **Select 3D-structures on this page to get Protein Data Base record(s) of crystal structure data of the nearest protein - this will yield 1IX5**
  - **Go to Structure (top of web page) and enter 1IX5 and click on its icon on the next page**
  - **To view a 3D-image of the protein, first download Cn3D from the NCBI site**

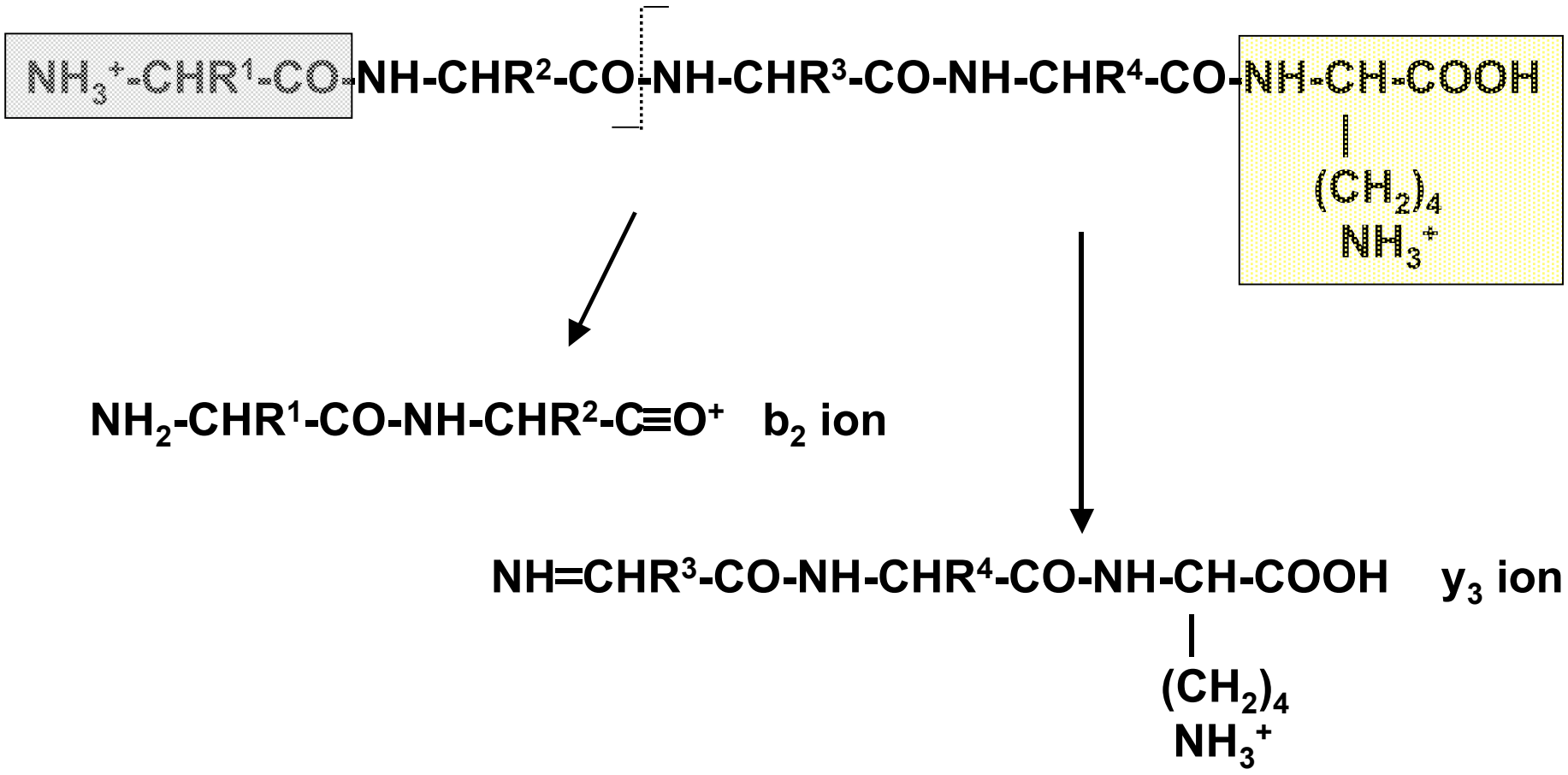
Examples for homework (due Feb 10th)

- **Identify the following proteins from these MALDI ions (corrected for isotope effects):**
  - **968.47, 1060.67, 1095.54, 1156.67, 1292.72, 2081.20 (human)**
  - **932.57, 1023.61, 1088.63, 1121.68, 1433.83, 1836.90 (rat)**
  - **937.60, 964.57, 1049.64, 1209.73, 1508.78, 1844.98 (rat)**
- **Set the number of tryptic cuts to 0 and try varying the mass accuracy from 0.1 to 1.0 Da. How does this alter the MOWSE score?**

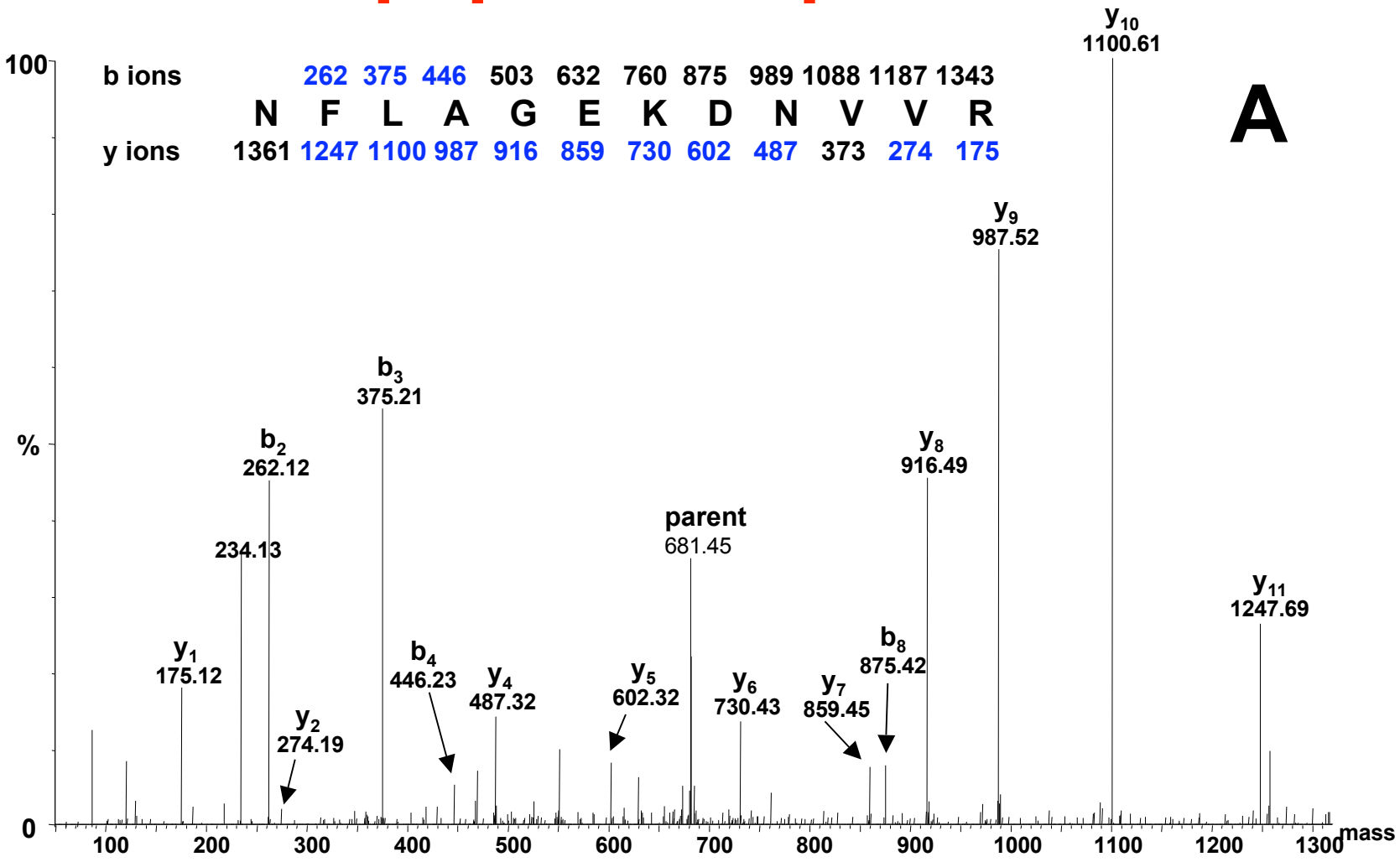
Sequencing of peptides

- **Using tandem mass spectrometry in a triple quadrupole, Q-tof, or ion trap instrument, the parent ion is first selected in the first quadrupole**
- **The parent ion is collided with argon gas and it breaks into fragments (daughter ions)**
- **By identifying the daughter ions, the peptide amino acid sequence is inferred**

Fragmentation of parent ion



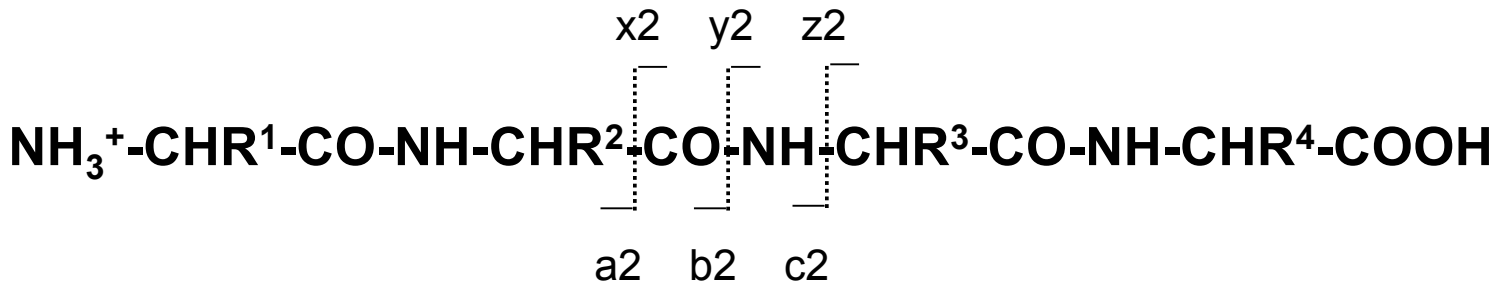
Identification of daughter ions and peptide sequence

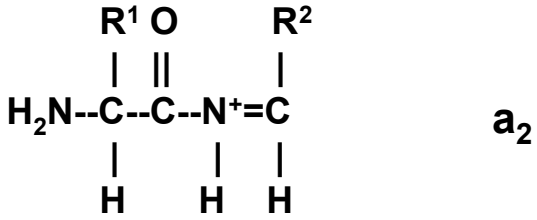


Amino acid residues masses

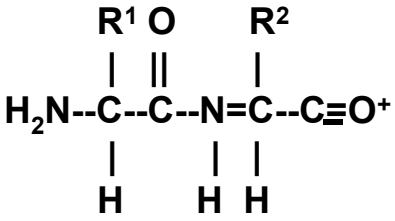
Alanine	71.037	Leucine	113.084
Arginine	156.101	Lysine	128.094
Asparagine	114.043	Methionine	131.040
Aspartic acid	115.027	Phenylalanine	147.068
Cysteine	103.009	Proline	97.053
Glutamic acid	129.043	Serine	87.032
Glutamine	128.058	Threonine	101.048
Glycine	57.021	Tryptophan	186.079
Histidine	137.059	Tyrosine	163.063
Isoleucine	113.084	Valine	99.068

Fragmenting a peptide

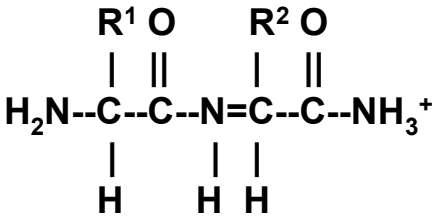




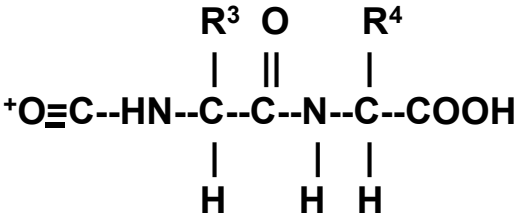
a_2



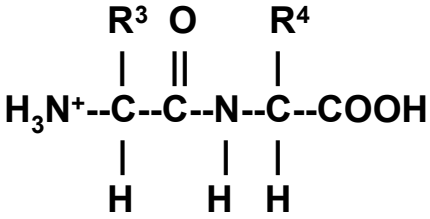
b_2



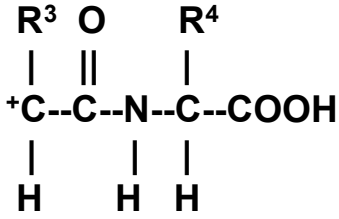
c_2



x_2



y_2

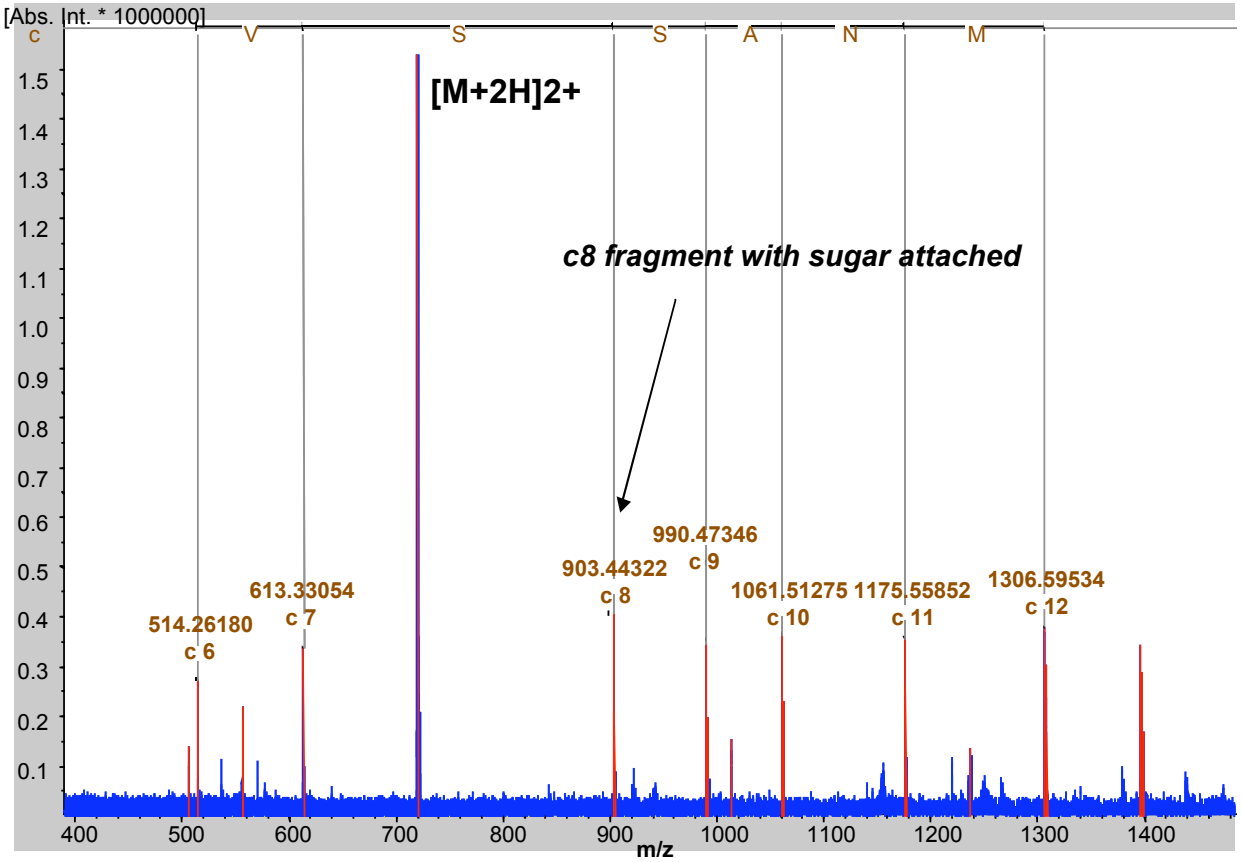


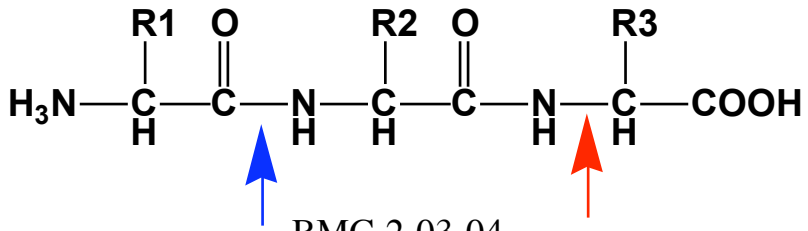
z_2

http://www.matrixscience.com/help/fragmentation_help.html

Sequencing O-GlcNAc peptides by ECD FT-ICR-MS

Casein kinase II - AGGSTPVSSANMMSG





b ion cleavage

BMG 2-03-04

c ion cleavage

Fragment ions of a small 5-mer peptide

Homework - write down the masses of the b and y ions

b_n = [residue masses + 1]

y_n = [residue masses + H_2O + 1]

A Q Y E K