

Mass Spectrometric Peptide Identification Using MASCOT

Dr. David Wishart

University of Alberta, Edmonton, Canada

david.wishart@ualberta.ca

MS Proteomics Applications

- **Protein identification/confirmation**
- **Protein sample purity determination**
- **Detection of post-translational modifications**
- **Detection of amino acid substitutions**
- **Determination of disulfide bonds (# & status)**
- **De novo peptide sequencing**
- **Monitoring protein folding (H/D exchange)**
- **Monitoring protein-ligand complexes/struct.**
- **3D Structure determination**

Protein Identification

- **2D-GE + MALDI-MS**
 - Peptide Mass Fingerprinting (PMF)
- **2D-GE + MS-MS**
 - MS Peptide Sequencing/Fragment Ion Searching
- **Multidimensional LC + MS-MS**
 - ICAT Methods (isotope labelling)
 - MudPIT (Multidimensional Protein Ident. Tech.)
- **1D-GE + LC + MS-MS**
- **De Novo Peptide Sequencing**

All require computers to process & analyze data

What is MASCOT?

- A (very) popular web-based tool from Matrix Science (www.matrixscience.com) for performing rapid, accurate, on-line MS analysis of peptides and proteins
- Supports 3 kinds of analyses
 - Peptide Mass Fingerprinting (PMF)
 - Sequence (tag) querying
 - MS/MS Ion searches

Matrix Science Website

Matrix Science - Home - Netscape

Matrix Science - Home

MATRIX SCIENCE

HOME : WHAT'S NEW : **MASCOT** : HELP : PRODUCTS : SUPPORT : CONTACT

Search Go

Home

Welcome

This site features **Mascot**, a powerful search engine that uses mass spectrometry data to identify proteins from primary sequence databases. To assist you, the [help text](#) for Mascot forms a substantial knowledge base concerning protein identification by MS.

If this is your first visit, please check for [browser compatibility](#) and read the [small print](#). If you include results from Mascot in a publication, please cite either this URL or Electrophoresis, **20(18)** 3551-67 (1999) ([abstract](#)).


We value your feedback and suggestions for new features. If you find any problems, errors, oversights, or just get unexpected results then please let us know.

For information on licensing Mascot for in-house use, please refer to our [Products](#) and [Support](#) pages. For recent news, check [What's New](#).

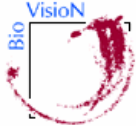
Matrix Science develops and markets software products which integrate mass spectrometry into bioinformatics. Our interests extend to all aspects of mass spectrometry in the life sciences. Please [contact us](#) to discuss:

- Developing new applications
- Consultancy in mass spectrometry and bioinformatics
- Systems analysis and integration


Collaborations

 Cancer Research Technology

Mascot incorporates code from [Mowse](#), developed by Darryl Pappin and David Perkins when working at the former Imperial Cancer Research Fund, and licensed from its technology transfer subsidiary, [Cancer Research Technology](#).

 BioVisioN

Matrix Science is collaborating with [BioVisioN](#) to develop improved data reduction software.

 LabVantage Solutions

[LabVantage Solutions](#) and Matrix Science are working together to develop data management and data mining solutions for proteomics.

start | My Documents | Matrix Science - Hom... | Microsoft PowerPoint ... | 9:41 AM

Mascot Home Page



Mascot Search

Mascot Help
Mascot Overview
Search parameter reference
Sequence databases
Data file format
Scoring algorithm
Results format
Error tolerant search
FAQ's
User Meeting Presentations
2004
More Help
Help Topic Index
Useful Links

- ◆ **Peptide Mass Fingerprint:** The experimental data are a list of peptide mass values from an enzymatic digest of a protein.
 - ◇ Example of results report
 - ◇ More information
- ◆ **Sequence Query:** One or more peptide mass values associated with information such as partial or ambiguous sequence strings, amino acid composition information, MS/MS fragment ion masses, etc. A super-set of a sequence tag query.
 - ◇ Example of results report
 - ◇ More information
- ◆ **MS/MS Ion Search:** Identification based on raw MS/MS data from one or more peptides.
 - ◇ Example of results report
 - ◇ More information

Search Form Defaults: Follow this link to save your preferred search form defaults as a browser cookie.

Why Mascot?

- Among the first to offer free web-based services for both PMF and MS/MS
- First to use probability-based scoring (PBS) or “Expect” values to rank matches and hits (significant improvement over all other scoring methods)
- Easy-to-use interface, fast, reliable, up-to-date databases, accurate – **a common industry standard**

Two Mascot Choices

- Matrix Science offers two choices for users:
- #1) A free, open access web-based system for occasional (1-10) queries per day **(this is what we'll use)**
- #2) A locally installed version for heavy use or high throughput MS and MS/MS labs (100's of queries/day)

Local Mascot Server

- License cost is ~\$4000 per CPU
- Single or dual processor Pentium 4, Xeon, Athlon, Opteron chips (300 MHz takes 200s/search, 3 GHz takes 20s)
- 2 Gbytes of RAM (key to performance)
- 120 Gbytes of Hard Disk (IDE) space to store all desired databases
- Can run on Windows or Linux (same)

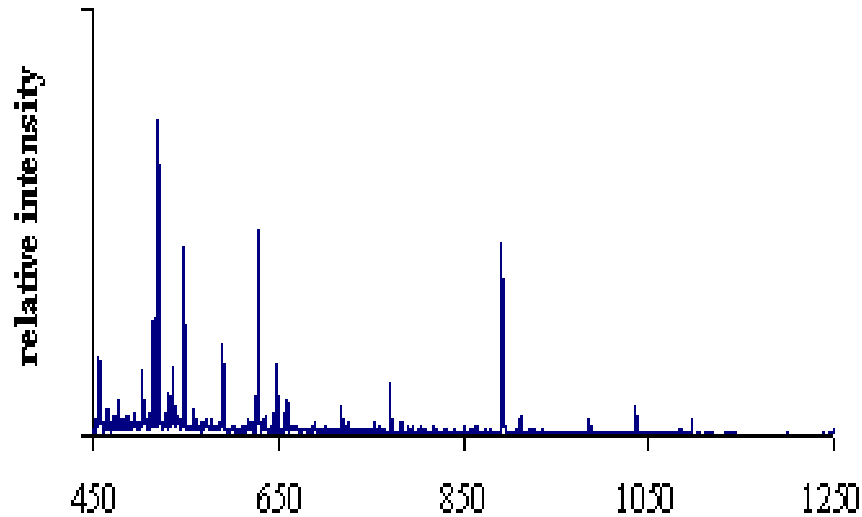
Local Mascot

- Allows you to customize your databases and to customize the frequency of database uploads
- **Mascot Distiller** – generates peak lists from just about any instrument (converts everything to a Mascot Generic File “GMF”)
- **Mascot Daemon** – allows you to do batch searches “press submit and go home” also allows monitoring of data flow on MS instrument and autoprocessing of that data

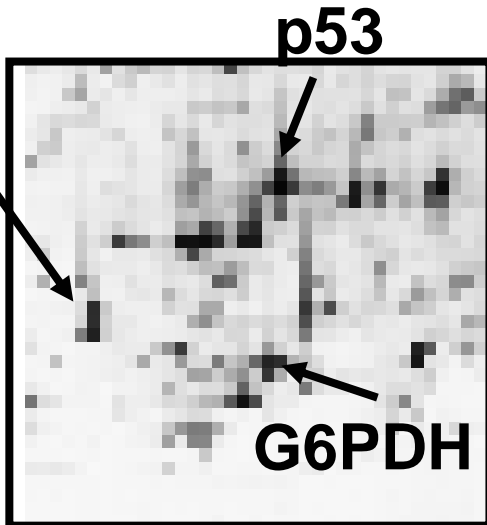
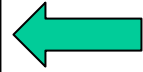
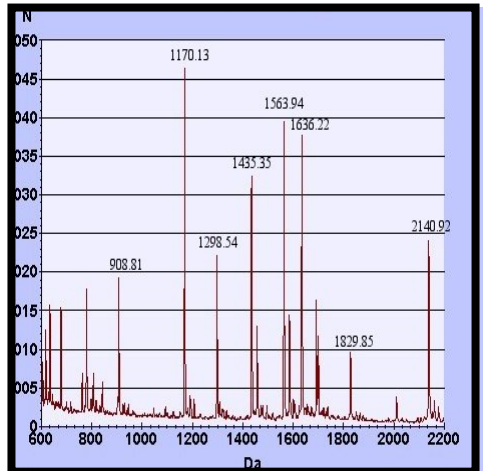
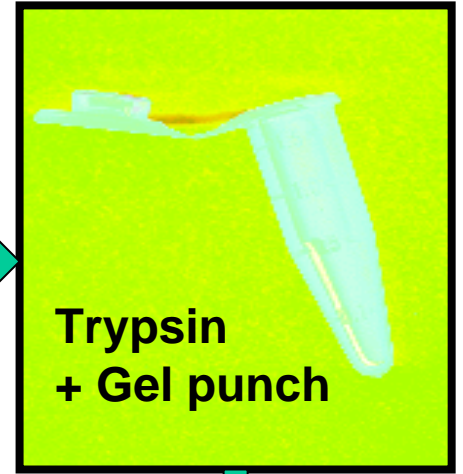
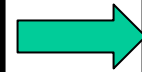
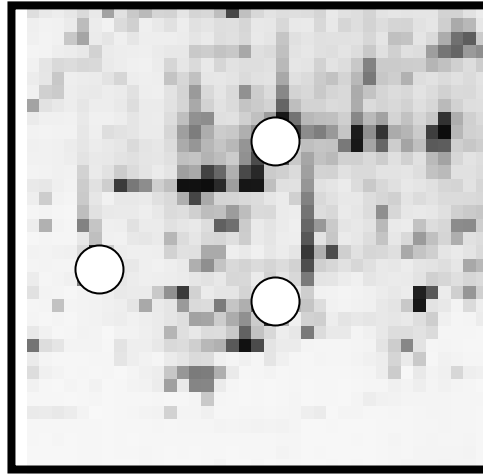
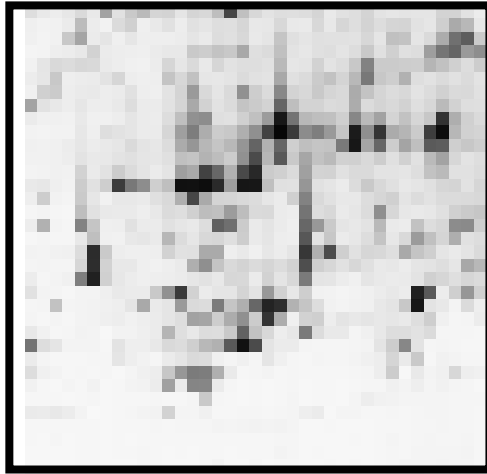
Mascot Databases & General Disk Needs

- Databases:
 - MSDB: 5 Gb
 - NCBIInr: 2 Gb
 - EST_others: 19 Gb
 - EST_human: 9 Gb
 - EST_mouse: 6 Gb
- Search results files
 - Range from 30 k to 300 Mb
 - e.g. 100 results / day @ 1 Mb each \approx 20 Gb / year
- Recommend at least 120 Gb drive.

Example #1 Peptide Mass Fingerprinting (PMF)



2D-GE + MALDI (PMF)



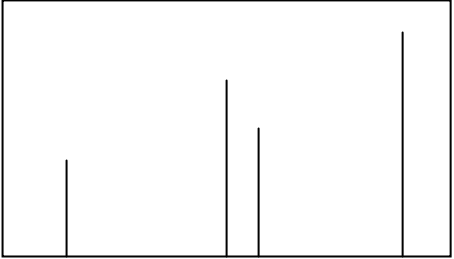
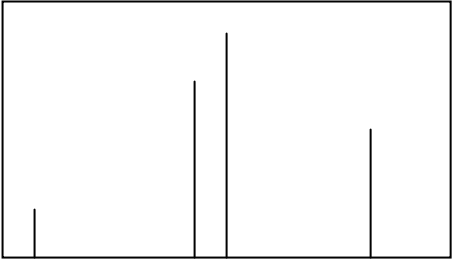
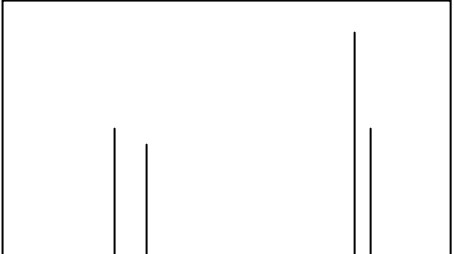
Peptide Mass Fingerprinting

- Used to identify protein spots on gels or protein peaks from an HPLC run
- Depends on the fact that if a peptide is cut up or fragmented in a known way, the resulting fragments (and resulting masses) are unique enough to identify the protein
- Requires a database of known sequences
- Uses software to compare observed masses with masses calculated from database

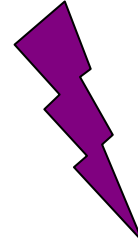
Principles of Fingerprinting

<u>Sequence</u>	<u>Mass (M+H)</u>	<u>Tryptic Fragments</u>
>Protein 1 acedfhsakdfgea sdfpkivtmeeewe ndadnfekqwfe	4842.05	acedfhsak dfgeasdfpk ivtmeeewendadnfek qwfe
>Protein 2 acekdfhsadfgea sdfpkivtmeeewe nkdadnfefqwfe	4842.05	acek dfhsadfgeasdfpk ivtmeeewenk dadnfefqwfe
>Protein 3 acedfhsadfgeka sdfpkivtmeeewe ndakdnfefqwfe	4842.05	acedfhsadfgek asdfpk ivtmeeewendak dnfefqwfe

Principles of Fingerprinting

<u>Sequence</u>	<u>Mass (M+H)</u>	<u>Mass Spectrum</u>
>Protein 1 acedfhsakdfgea sdfpkivtmeeewe ndadnfekqwfe	4842.05	
>Protein 2 acekdfhsadfgea sdfpkivtmeeewe nkdadnfefqwfe	4842.05	
>Protein 3 acedfhsadfgeka sdfpkivtmeeewe ndakdnfefqwfe	4842.05	

Protease Cleavage Rules



Trypsin

XXX[KR]--[!P]XXX

Chymotrypsin

XX[FYW]--[!P]XXX

Lys C

XXXXXK-- XXXXX

Asp N endo

XXXXXD-- XXXXX

CNBr

XXXXXM--XXXXX

Why Trypsin?

- Robust, stable enzyme
- Works over a range of pH values & Temp.
- Quite specific and consistent in cleavage
- Cuts frequently to produce “ideal” MW peptides
- Inexpensive, easily available/purified
- Does produce “autolysis” peaks (which can be used in MS calibrations)
 - 1045.56, 1106.03, 1126.03, 1940.94, 2211.10, 2225.12, 2283.18, 2299.18

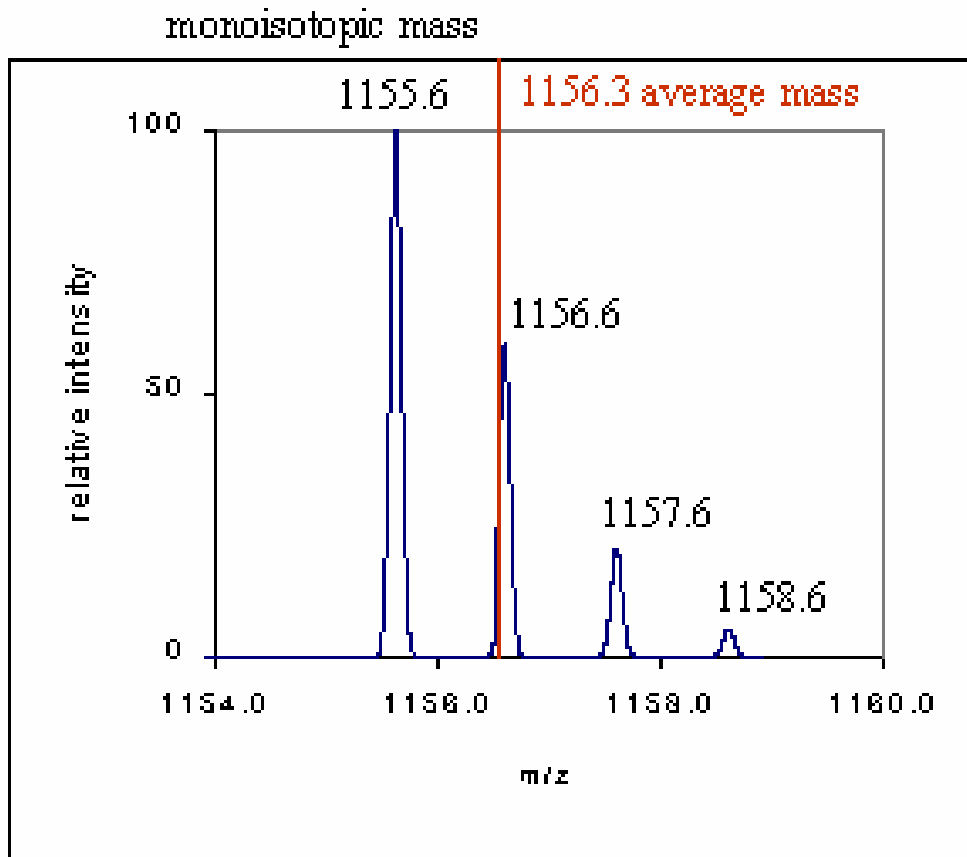
Calculating Peptide Masses

- Sum the monoisotopic residue masses
- Add mass of H₂O (18.01056)
- Add mass of H⁺ (1.00785 to get M+H)
- If Met is oxidized add 15.99491
- If Cys has acrylamide adduct add 71.0371
- If Cys is iodoacetylated add 58.0071
- Other modifications are listed at
 - <http://www.abrf.org/index.cfm/dm.home>
- Only consider peptides with masses > 400

Post-Translational Modifications (PTM)

	abbreviation	monoisotopic	average				
Acetylation	ACET	42.0106	42.0373	Geranyl-geranyl	GERA	272.2504	272.4741
Alkylation	ALKY	14.01564	14.02688	Gamma-carboxyglutamic acid	GGLU	43.98983	44.0098
Amidation	AMID	-0.9840	-0.9847	O-GlcNAc	GLCN	203.0794	203.1950
Beta-methylthiolation	BMTH	45.9877118	46.08688	Glucosylation (Glycation)	GLUC	162.0528	162.1424
Biotin	BIOT	226.0776	226.2934	Glutathionylation	GLUT	305.0680814	305.3056
Bromination	BROM	77.9105	78.9	Hydroxylation	HYDR	15.9949	15.9994
Carbamylation	CAM	43.00581	43.02502	Lipoyl	LIPY	188.033	188.3027
Citrullination	CTTR	0.9840276	0.98476	Methylation	METH	14.0157	14.0269
C-Mannosylation	CMAN	162.052823	162.1424	Myristoylation	MYRI	210.1984	210.3598
Cysteine sulfenic acid (-SOH)	CSEA	15.9949146	15.9994	S-Nitrosylation	NTRY	28.99017	28.99816
Cysteine sulfinic acid (-SO₂H)	CSIA	31.9898292	31.9988	N-Octanoate	OCTA	126.1044	126.1986
Deamidation	DEAM	0.9840	0.9847	Palmitoylation	PALM	238.2297	238.4136
N-acyl diglyceride cysteine (tripalmitate)	DIAC	788.7258	789.3202	Phosphorylation	PHOS	79.9663	79.9799
Dimethylation	DIMETH	28.0314	28.0538	Pyridoxal phosphate	PLP	229.014	229.129
FAD	FAD	783.1415	783.542	Phosphopantetheine	PPAN	339.078	339.3234
Farnesylation	FARN	204.1878	204.3556	Pyrrolidone carboxylic acid	PYRR	-17.0266	-17.0306
Formylation	FORM	27.9949	28.0104	Sulfation	SULF	79.9568	80.0642
				Trimethylation	TRIMETH	42.0471	42.0807

Masses in MS



- **Monoisotopic mass is the mass determined using the masses of the most abundant isotopes**
- **Average mass is the abundance weighted mass of all isotopic components**

Amino Acid Residue Masses

Monoisotopic Mass

Glycine	57.02147	Aspartic acid	115.02695
Alanine	71.03712	Glutamine	128.05858
Serine	87.03203	Lysine	128.09497
Proline	97.05277	Glutamic acid	129.04264
Valine	99.06842	Methionine	131.04049
Threonine	101.04768	Histidine	137.05891
Cysteine	103.00919	Phenylalanine	147.06842
Isoleucine	113.08407	Arginine	156.10112
Leucine	113.08407	Tyrosine	163.06333
Asparagine	114.04293	Tryptophan	186.07932

Amino Acid Residue Masses

Average Mass

Glycine	57.0520	Aspartic acid	115.0886
Alanine	71.0788	Glutamine	128.1308
Serine	87.0782	Lysine	128.1742
Proline	97.1167	Glutamic acid	129.1155
Valine	99.1326	Methionine	131.1986
Threonine	101.1051	Histidine	137.1412
Cysteine	103.1448	Phenylalanine	147.1766
Isoleucine	113.1595	Arginine	156.1876
Leucine	113.1595	Tyrosine	163.1760
Asparagine	114.1039	Tryptophan	186.2133

Preparing a Peptide Mass Fingerprint Database

- Take a protein sequence database (Swiss-Prot or nr-GenBank)
- Determine cleavage sites and identify resulting peptides for each protein entry
- Calculate the mass ($M+H$) for each peptide
- Sort the masses from lowest to highest
- Have a pointer for each calculated mass to each protein accession number in databank

Building A PMF Database

Sequence DB

>P12345
acedfhsakdfqea
sdfpkivtmeeewe
ndadnfekqwfe

>P21234
acekdfhsadfqea
sdfpkivtmeeewe
nkdadnfefqwfe

>P89212
acedfhsadfgeka
sdfpkivtmeeewe
ndakdnfefqwfe

Calc. Tryptic Frags

acedfhsak
dfgeasdfpk
ivtmeeewendadnfek
gwfe

acek
dfhsadfgeasdfpk
ivtmeeewenk
dadnfefqwfe

acedfhsadfgek
asdfpk
ivtmeeewendak
dnfefqwfe

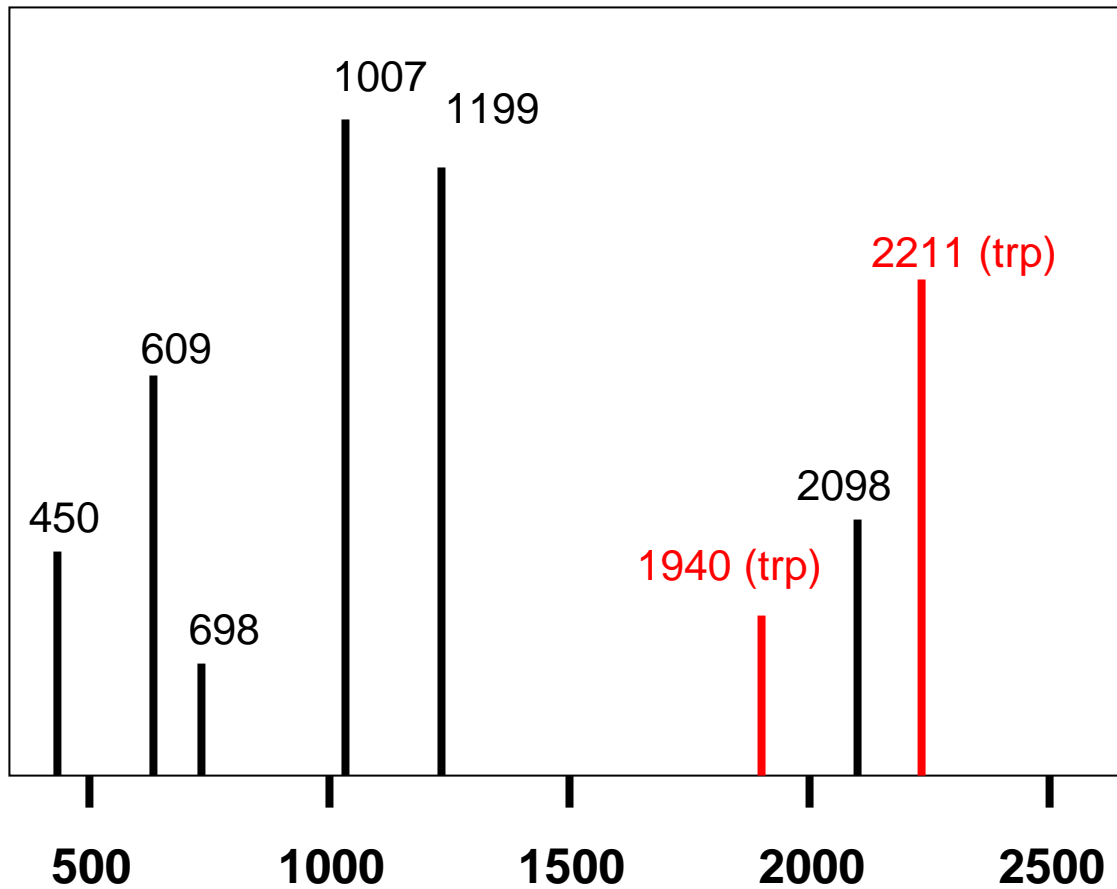
Mass List

450.2017 (P21234)
609.2667 (P12345)
664.3300 (P89212)
1007.4251 (P12345)
1114.4416 (P89212)
1183.5266 (P12345)
1300.5116 (P21234)
1407.6462 (P21234)
1526.6211 (P89212)
1593.7101 (P89212)
1740.7501 (P21234)
2098.8909 (P12345)

The Fingerprint (PMF) Algorithm

- Take a mass spectrum of a trypsin-cleaved protein (from gel or HPLC peak)
- Identify as many masses as possible in spectrum (avoid autolysis peaks)
- Compare query masses with database masses and calculate # of matches or matching score (based on length and mass difference)
- Rank hits and return top scoring entry – this is the protein of interest

Query (MALDI) Spectrum




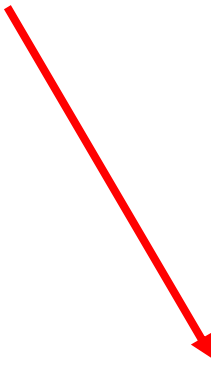


Query vs. Database

Query Masses

Database Mass List

Results

450.2201		450.2017 (P21234)	2 Unknown masses
609.3667		609.2667 (P12345)	
698.3100		664.3300 (P89212)	
1007.5391		1007.4251 (P12345)	1 hit on P21234
1199.4916		1114.4416 (P89212)	3 hits on P12345
2098.9909		1183.5266 (P12345)	
		1300.5116 (P21234)	
		1407.6462 (P21234)	
		1526.6211 (P89212)	
		1593.7101 (P89212)	
		1740.7501 (P21234)	
		2098.8909 (P12345)	

Conclude the query protein is P12345

What You Need To Do PMF

- A list of query masses (as many as possible)
- Protease(s) used or cleavage reagents
- Databases to search (SWProt, NR, Organism)
- Estimated mass and pI of protein spot (**opt**)
- Cysteine (or other) modifications
- Minimum number of hits for significance
- Mass tolerance (100 ppm = 1000.0 ± 0.1 Da)
- *A PMF website (Prowl, ProFound, Mascot, etc.)*

PMF on the Web

- **Mascot**
 - www.matrixscience.com
- **ProFound**
 - http://129.85.19.192/profound_bin/WebProFound.exe
- **MOWSE**
 - <http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse>
- **PeptideSearch**
 - <http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html>
- **Peptident**
 - <http://us.expasy.org/tools/peptident.html>

Mascot – PMF Query



Mascot Search

Mascot Help
Mascot Overview
Search parameter reference
Sequence databases
Data file format
Scoring algorithm
Results format
Error tolerant search
FAQ's
User Meeting Presentations
2004
More Help
Help Topic Index
Useful Links

- ◆ **Peptide Mass Fingerprint:** The experimental data are a list of peptide mass values from an enzymatic digest of a protein.
 - ◊ Example of results report
 - ◊ More information
- ◆ **Sequence Query:** One or more peptide mass values associated with information such as partial or ambiguous sequence strings, amino acid composition information, MS/MS fragment ion masses, etc. A super-set of a sequence tag query.
 - ◊ Example of results report
 - ◊ More information
- ◆ **MS/MS Ion Search:** Identification based on raw MS/MS data from one or more peptides.
 - ◊ Example of results report
 - ◊ More information

Search Form Defaults: Follow this link to save your preferred search form defaults as a browser cookie.

click

MASCOT Peptide Mass Fingerprint

Your name	<input type="text"/>	Email	<input type="text"/>
Search title	<input type="text"/>		
Database	MSDB <input type="button" value="v"/>		
Taxonomy	All entries <input type="button" value="v"/>		
Enzyme	Trypsin <input type="button" value="v"/>	Allow up to	1 <input type="button" value="v"/> missed cleavages
Fixed modifications	<input type="text" value="AB_old_ICATd0 (C)"/> <input type="button" value="▲"/> <input type="text" value="AB_old_ICATd8 (C)"/> <input type="button" value="▲"/> <input type="text" value="Acetyl (K)"/> <input type="button" value="▲"/> <input type="text" value="Acetyl (N-term)"/> <input type="button" value="▲"/> <input type="text" value="Amide (C-term)"/> <input type="button" value="▼"/>	Variable modifications	<input type="text" value="AB_old_ICATd0 (C)"/> <input type="button" value="▲"/> <input type="text" value="AB_old_ICATd8 (C)"/> <input type="button" value="▲"/> <input type="text" value="Acetyl (K)"/> <input type="button" value="▲"/> <input type="text" value="Acetyl (N-term)"/> <input type="button" value="▲"/> <input type="text" value="Amide (C-term)"/> <input type="button" value="▼"/>
Protein mass	<input type="text"/> kDa	Peptide tol. ±	<input type="text" value="1.0"/> Da <input type="button" value="v"/>
Mass values	<input checked="" type="radio"/> MH ⁺ <input type="radio"/> M _r	Monoisotopic	<input checked="" type="radio"/> Average <input type="radio"/>
Data file	<input type="text"/> <input type="button" value="Browse..."/>		
Query NB Contents of this field are ignored if a data file is specified.	<input type="text"/>		
Overview	<input type="checkbox"/>	Report top	20 <input type="button" value="v"/> hits
<input type="button" value="Start Search ..."/>		<input type="button" value="Reset Form"/>	

Exercise #1

- **Analysis of a yeast protein (75 KDa) treated with iodoacetamide, trypsinized and subject to MALDI-TOF**
- **Go to “Worked Example 1” in your notes to follow instructions**
- **Access your PMF data at:**
<http://gchelpdesk.ualberta.ca/ABRF2005/>

listed as Example1.txt

What Are Missed Cleavages?

Sequence

>Protein 1
acedfhsakdfgea
sdfpkivtmeeewe
ndadnfekqwfe

Tryptic Fragments (no missed cleavage)

acedfhsak (1007.4251)
dfgeasdfpk (1183.5266)
ivtmeeewendadnfek (2098.8909)
qwfe (609.2667)

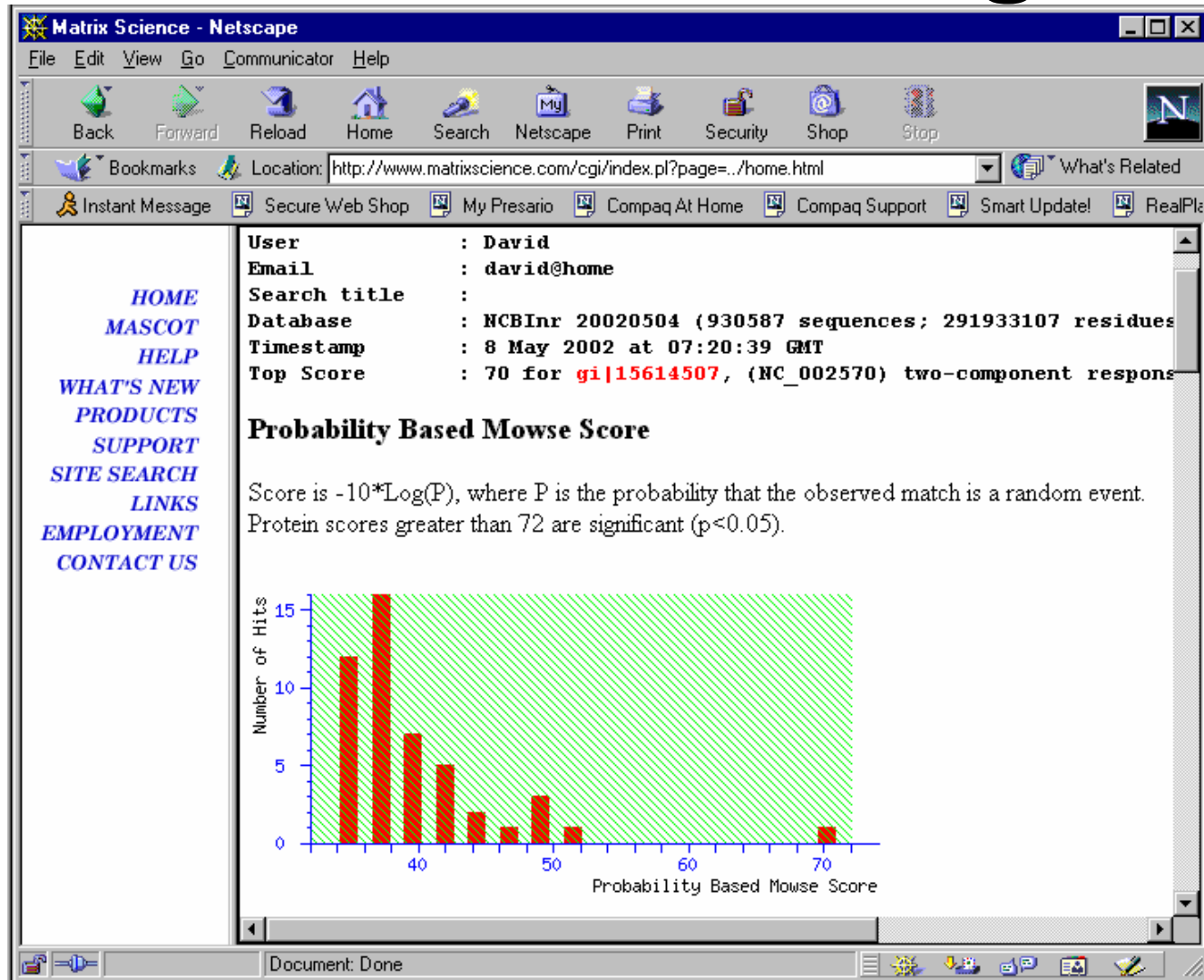
Tryptic Fragments (1 missed cleavage)

acedfhsak (1007.4251)
dfgeasdfpk (1183.5266)
ivtmeeewendadnfek 2098.8909)
qwfe (609.2667)
acedfhsakdfgeasdfpk (2171.9338)
ivtmeeewendadnfekqwfe (2689.1398)
dfgeasdfpkivtmeeewendadnfek (3263.2997)

Mascot Databases

Database	Comment
EST	EST divisions of Genbank, (currently EST_human, EST_mouse, EST_others)
MSDB	Comprehensive, non-identical protein database
NCBItr	Comprehensive, non-identical protein database
OWL	Non-identical protein database (obsolete)
Random	Random sequences for verifying scoring statistics
SwissProt	High quality, curated protein database

MASCOT Scoring



Why Probability-Based Scoring?

- Will explain PBS later...
- Offers a simple numerical (and graphical) assessment of whether a result is significant
- More reliable/accurate than simple mass or peptide cut-off techniques
- Allows both MS and PMF data to be scored the same way
- Scores from different searches or different databases can be easily & directly compared

Mascot Scoring

- The statistics of peptide fragment matching in MS (or PMF) is very similar to the statistics used in BLAST
- The scoring probability appears to follow an extreme value distribution
- High scoring segment pairs (in BLAST) are analogous to high scoring mass matches in Mascot
- Mascot scoring system is based on the MOWSE scoring system

MOWSE

- **MO**lecular **W**eight **S**Earch
- **Scoring system based on peptide frequency distribution from the OWL non redundant protein Database**

Pappin DJC, Hojrup P, and Bleasby AJ (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3**:327-332

MOWSE

Sequence

>Protein 1

acedfhsakdfgea
sdfpkivtmeeewe
ndadnfekqwfe

>Protein 2

acekdfhsadfgea
sdfpkivtmeeewe
nkdadnfefqwfe

>Protein 3

MASMGTLAFD EYGRPFLLIK
DQDRKSRLMG LEALKSHIM
A AKAVANTMRT SLGPNGLD
KMMVDKDGDTV TNDGAT
ILSM MDVDHQIAKL MVELS
KSQDD EIGDGTTGVV VLAG
ALLEEAEQLLD RGIHP IRIAD

Mass (M+H)

4842.05

4842.05

14563.36

Tryptic Fragments

acedfhsak
dfgeasdfpk
ivtmeeewendadnfek
gwfe

acek
dfhsadfgeasdfpk
ivtmeeewenk
dadnfefqwfe

SQDDEIGDGTTGVVVLGALLEEAEQLLDR2
DGDVTVTNDGATILSMMDVD HQIAK
MASMGTLAFDEYGRPFLLIK2
TSLGPNGLDK
LMGLEALK
LMVELSK
AVANTMR
SHIMAAK
GIHPIR
MMVDK
DQDR

MOWSE

1. Group Proteins into 10 kDa 'bins'.

0-10 kDa

```
>Protein 1  
acedfhsakdfqea  
sdfpkivtmeeewe  
ndadnfekqwfel
```

4954.13

```
>Protein 2  
acekdfhsadfqea  
sdfpkivtmeeewe  
nkdadnfekqwfekq  
wfei
```

5672.48

10-20 kDa

```
>Protein 3  
MASMGTLAFD EYGRPFLIK  
DQDRKSRLMG LEALKSHIM  
A AKAVANTMRT SLGPNGLD  
KMMVDKDGDTV TNDGAT  
ILSM MDVDHQIAKL MVELS  
KSQDD EIGDGTGVV VLAG  
ALLEEAEQLLDRGIHP IRIAD
```

14563.36

MOWSE

2. For each protein, place fragments into 100 Da bins.

>Protein 1
 acedfhsakdfgea
 sdfpkivtmeeewe
 ndadnfekqwfel

<u>Mol. Wt.</u>	<u>Fragment</u>
2098.8909	IVTMEEEWENDADNFEK
1183.5266	DFQEASDFPK
1007.4251	ACEDFHSAK
722.3508	QWFEL

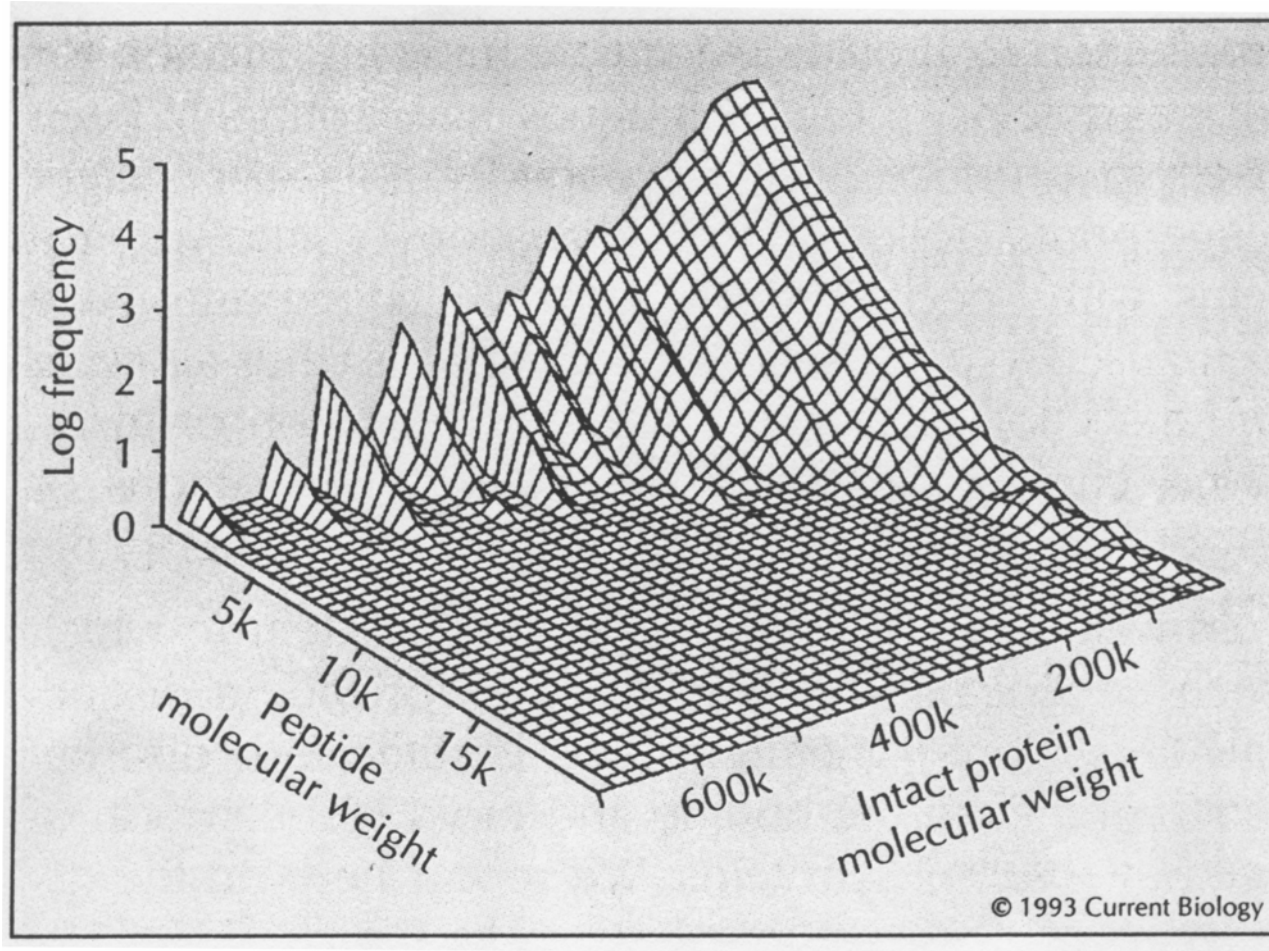
>Protein 2
 acekdfhsadfgea
 sdfpkivtmeeewe
 nkdadnfefqwfekq
 wfei

1740.7500	DFHSADFQEASDFPK
1407.6460	IVTMEEEWENK
1456.6127	DADNFEQWFEK
722.3508	QWFEI

<u>Bin</u>	<u>Fragment</u>
2000-2100	IVTMEEEWENDADNFEK
1900-2000	
1800-1900	
1700-1800	DFHSADFQEASDFPK
1600-1700	
1500-1600	
1400-1500	IVTMEEEWENK, DADNFEQWFE
1300-1400	
1200-1300	
1100-1200	DFQEASDFPK
1000-1100	ACEDFHSAK
900-1000	
800-900	
700-800	
600-700	QWFEL, QWFEI
500-600	
400-500	

MOWSE

The MOWSE frequency distribution plot looks like this:



MOWSE

3. Divide the number of fragments for each bin by the total number of fragments for each 10 kDa protein interval

<u>Bin</u>	<u>Fragment</u>	<u>Total</u>	<u>Frequency</u>
2000-2100	IVTMEEEWENDADNFEK	1	0.125
1900-2000		0	0.000
1800-1900		0	0.000
1700-1800	DFHSADFQEASDFPK	1	0.125
1600-1700		0	0.000
1500-1600		0	0.000
1400-1500	IVTMEEEWENK, DADNFEQWFE	2	0.250
1300-1400		0	0.000
1200-1300		0	0.000
1100-1200	DFQEASDFPK	1	0.125
1000-1100	ACEDFHSAK	1	0.125
900-1000		0	0.000
800-900		0	0.000
700-800		0	0.000
600-700	QWFEL, QWFEI	2	0.250
500-600		0	0.000
400-500		0	0.000

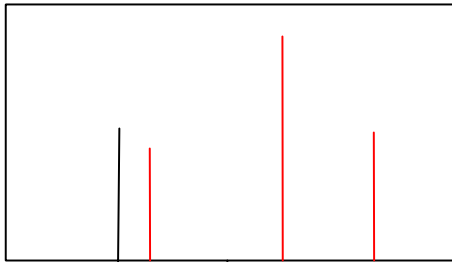
MOWSE

4. For each 10 kD interval, normalize to the largest bin value

<u>Bin</u>	<u>Fragment</u>	<u>Total</u>	<u>Frequency</u>	<u>Normalized</u>
2000-2100	IVTMEEEWENDADNFEK	1	0.125	0.5
1900-2000		0	0.000	0
1800-1900		0	0.000	0
1700-1800	DFHSADFQEASDFPK	1	0.125	0.5
1600-1700		0	0.000	0
1500-1600		0	0.000	0
1400-1500	IVTMEEEWENK, DADNFEQWFE	2	0.250	1
1300-1400		0	0.000	0
1200-1300		0	0.000	0
1100-1200	DFQEASDFPK	1	0.125	0.5
1000-1100	ACEDFHSK	1	0.125	0.5
900-1000		0	0.000	0
800-900		0	0.000	0
700-800		0	0.000	0
600-700	QWFEL, QWFEI	2	0.250	1
500-600		0	0.000	0
400-500		0	0.000	0

MOWSE

5. Compare spectrum masses against fragment mass list for each protein in the database. Retrieve the frequency score for each match and multiply.



1740.7500
1456.6127
722.3508

<u>Bin</u>	<u>Fragment</u>	<u>Total</u>	<u>Frequency</u>	<u>Normalized</u>
2000-2100	IVTMEEEWENDADNFEK	1	0.125	0.5
1900-2000		0	0.000	0
1800-1900		0	0.000	0
1700-1800	DFHSADFQEASDFPK	1	0.125	0.5
1600-1700		0	0.000	0
1500-1600		0	0.000	0
1400-1500	IVTMEEEWENK, DADNFEQWFE	2	0.250	1
1300-1400		0	0.000	0
1200-1300		0	0.000	0
1100-1200	DFQEASDFPK	1	0.125	0.5
1000-1100	ACEDFHSAK	1	0.125	0.5
900-1000		0	0.000	0
800-900		0	0.000	0
700-800		0	0.000	0
600-700	QWFEL, QWFEI	2	0.250	1
500-600		0	0.000	0
400-500		0	0.000	0

$$0.5 \times 1 \times 1 = 0.5$$

MOWSE

6. Invert and multiply, and normalize to an 'average' protein of 50 000 k Da:

$$P_N = \text{product of distribution frequency scores} \\ = 0.5 \times 1 \times 1 = 0.5$$


$$\text{Score} = \frac{50\,000}{P_N \times H} \quad \begin{array}{l} H = \text{'Hit' Protein MW} \\ = 5672.48 \end{array}$$

$$= \frac{50\,000}{0.5 \times 5672.48} = 17.62$$

MOWSE

- 😊 Takes into account relative abundance of peptides in the database when calculating scores
- 😊 Protein size is compensated for
- 😞 The model consists of numerous spaces separated by 100 Da (the average aa mass)
- 😞 Does not provide a measure of confidence for the prediction

MASCOT

- **Probability-based MOWSE scoring**
- **The probability that the observed match between experimental data and a protein sequence is a random event is approximately calculated for each protein in the sequence database.**
- ** Probability model details not published**

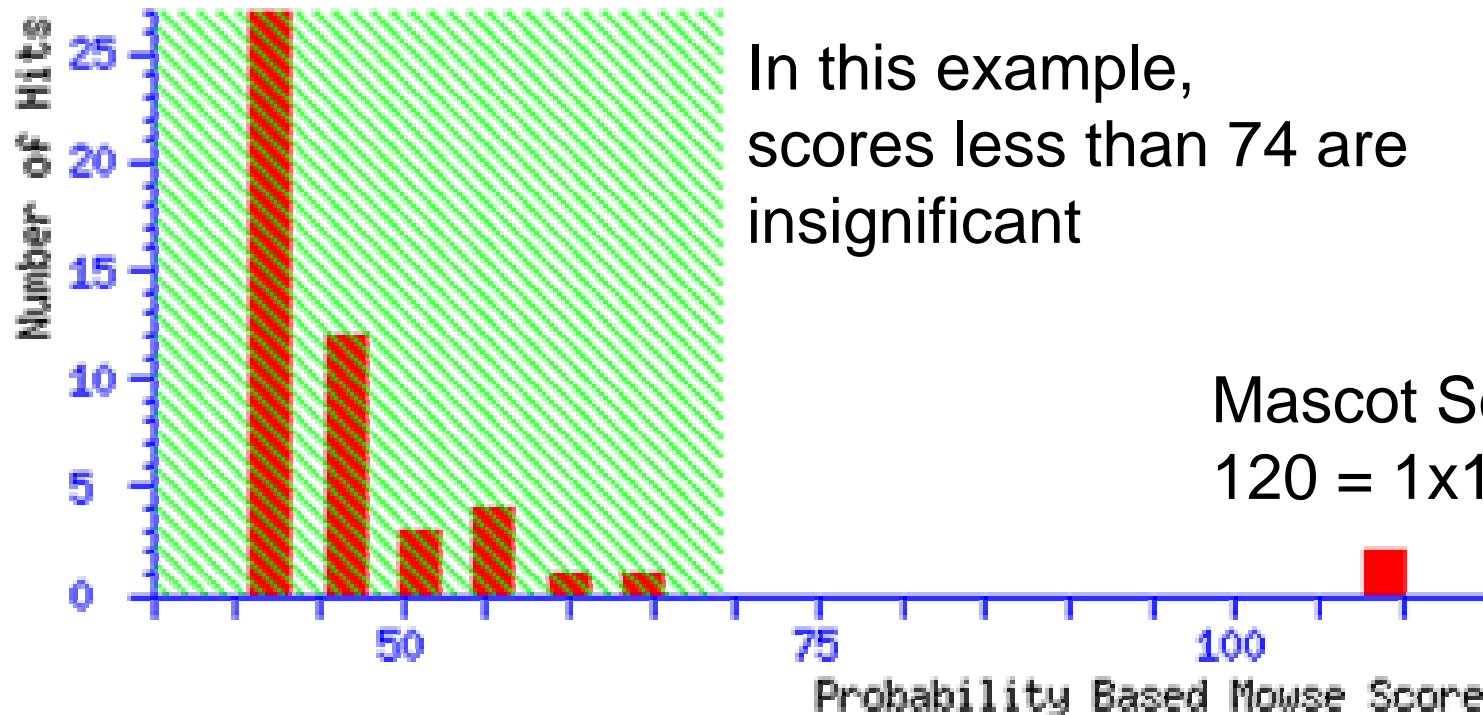
Perkins DN, Pappin DJC, Creasy DM, and Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**:3551-3567.

Mascot/Mowse Scoring

- The Mascot Score is given as $S = -10 \cdot \log(P)$, where P is the probability that the observed match is a random event
- Try to aim for probabilities where $P < 0.05$ (less than a 5% chance the peptide mass match is random)
- With today's databases, Mascot scores greater than 76 are significant ($p < 0.05$)
- We show in the Mascot Lab that a score's statistical significance is a complex function of database size, mass window tolerance, etc.

Mascot Scoring

- The Mascot Score is given as $S = -10 \cdot \log(P)$, where P is the probability that observed match is a random event
- The significance of that result depends on the size of the database being searched. Mascot shades in green the insignificant hits using a $P=0.05$ cutoff



Advantages of PMF

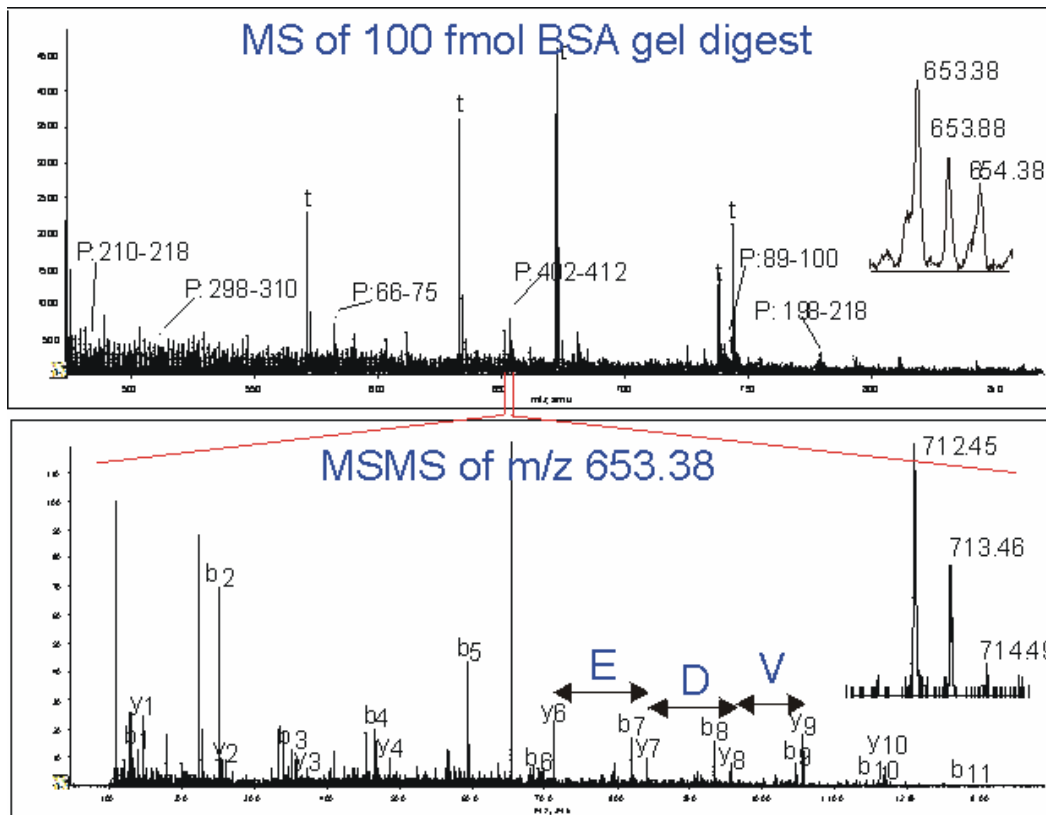
- Uses a “robust” & inexpensive form of MS (MALDI)
- Doesn't require too much sample optimization
- Can be done by a moderately skilled operator (don't need to be an MS expert)
- Widely supported by web servers
- Improves as DB's get larger & instrumentation gets better
- *Very amenable to high throughput robotics (up to 500 samples a day)*

Limitations With PMF

- Requires that the protein of interest already be in a sequence database
- Spurious or missing critical mass peaks always lead to problems
- Mass resolution/accuracy is critical, best to have <20 ppm mass resolution
- Generally found to only be about 40% effective in positively identifying gel spots

Example #2 MS/MS

Identification of a Protein from a Peptide Mixture



MS-MS for Protein ID

- Proteins are isolated (from gel or HPLC) and subjected to *tryptic* digestion
- Peptides are sent through ionizer and into a collision cell where the *doubly charged* ions are selected and fragmented through collision induced decay (CID)
- The resulting singly charged ions (daughter ions) are analyzed to determine the sequence or to ID the parent peptide

Why Trypsin for MS-MS?

- CID of peptides less than 2-3 kD is most reliable for MS-MS studies – The frequency of tryptic cleavage guarantees that most peptides will be of this size
- Trypsin cleaves on the C-terminal side of arginine and lysine. By putting the basic residues at the C-terminus, peptides fragment in a more predictable manner throughout the length of the peptide

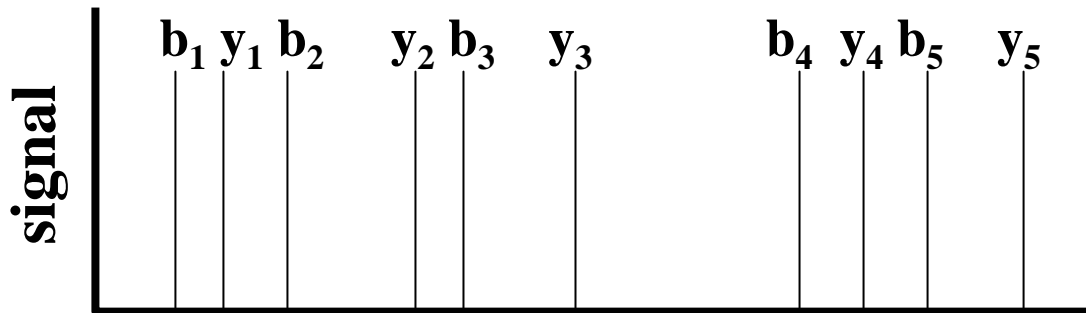
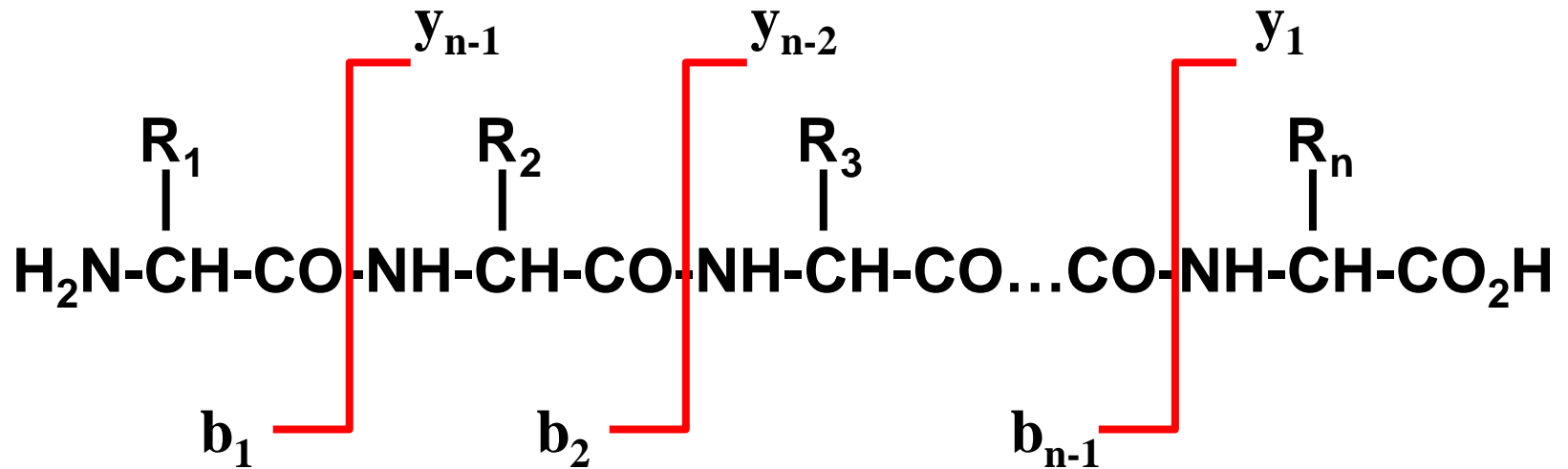
Why Double Charges?

- **Easiest spectra to interpret are those obtained from doubly-charged peptide precursors, where the resulting fragment ions are mostly singly-charged**
- **Doubly-charged precursors also fragment such that most of the peptide bonds break with comparable frequency, such that one is more likely to derive a complete sequence**

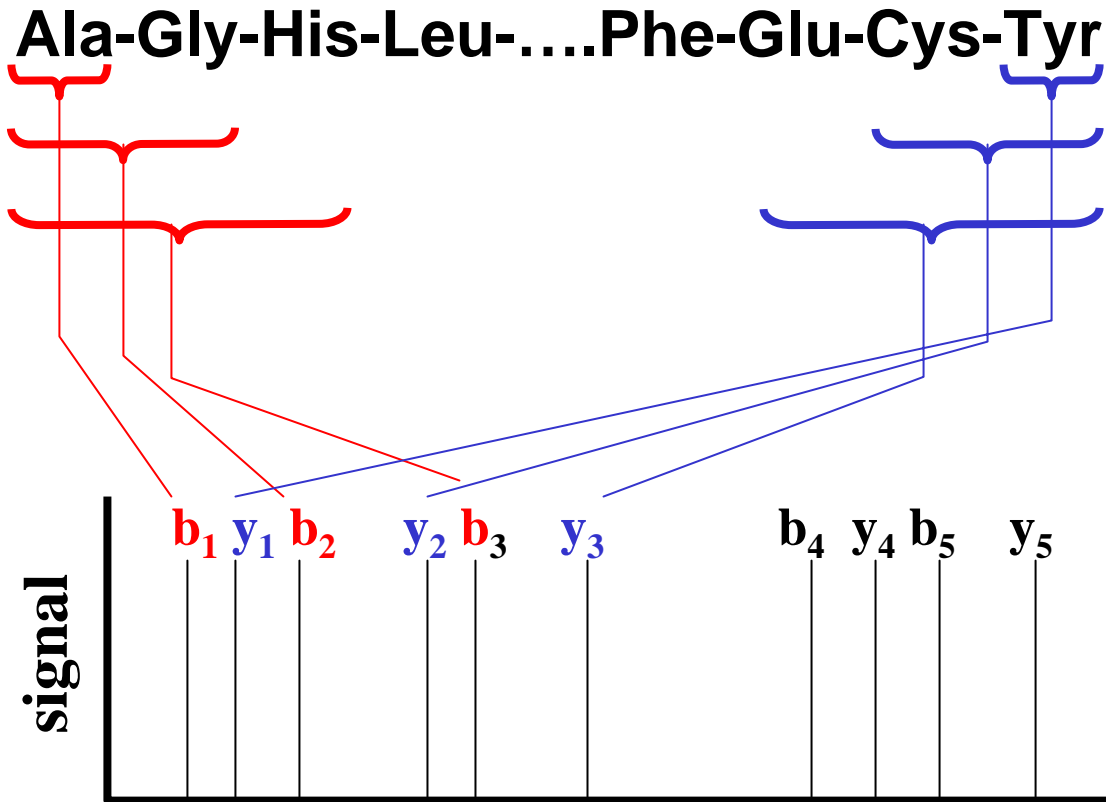
MS-MS & Peptide Fragments

- When peptides or proteins are admitted to a collision cell the peptide usually fragments at the weakest bond (the peptide bond, but some CH-NH and CH-CO breakage also occurs)
- Collision conditions have to be optimized for each peptide
- Two main types of daughter ions are produced -- “b” ions and “y” ions

MS-MS Peptide Fragmentation



MS-MS Peptide Fragmentation



Different MS-MS Instruments Yield Different Spectra

- A typical QTOF or triple quad MS-MS spectrum of a tryptic peptide contains a continuous series of y-type ions. The b-type ions are usually seen only at lower masses below the precursor m/z value
- Ion trap CID data of tryptic peptides is different in that one often finds a continuous series of both b-type and y-type ions throughout the spectrum

MS/MS – The Movie (Kathleen Binns)

- <http://www.mshri.on.ca/pawson/ms/movie.html>

Protein ID by MS-MS

- Peptide fragments from target protein are sequenced by MS-MS using a variety of algorithms (SEQUEST, Mascot) or via manual methods
- The peptide fragment sequences are sent to BLAST to be queried against a protein sequence database
- The protein having the highest number of sequence matches is ID'd as the target

MS-MS & Proteomics

Advantages

- Provides precise sequence-specific data
- More informative than PMF methods (>90%)
- Can be used for de-novo sequencing (not entirely dependent on databases)
- Can be used to ID post-trans. modifications

Disadvantages

- Requires more handling, refinement and sample manipulation
- Requires more expensive and complicated equipment
- Requires high level expertise
- Slower, not generally high throughput

Mascot – MS/MS Query



Mascot Search

Mascot Help
Mascot Overview
Search parameter reference
Sequence databases
Data file format
Scoring algorithm
Results format
Error tolerant search
FAQ's
User Meeting Presentations
2004
More Help
Help Topic Index
Useful Links

- ◆ **Peptide Mass Fingerprint:** The experimental data are a list of peptide mass values from an enzymatic digest of a protein.
 - ◇ Example of results report
 - ◇ More information
- ◆ **Sequence Query:** One or more peptide mass values associated with information such as partial or ambiguous sequence strings, amino acid composition information, MS/MS fragment ion masses, etc. A super-set of a sequence tag query.
 - ◇ Example of results report
 - ◇ More information
- ◆ **MS/MS Ion Search:** Identification based on raw MS/MS data from one or more peptides.
 - ◇ Example of results report
 - ◇ More information

Search Form Defaults: Follow this [link](#) to save your preferred search form defaults as a browser cookie.

click



MASCOT MS/MS Ions Search

Your name	<input type="text"/>	Email	<input type="text"/>
Search title	<input type="text"/>		
Database	MSDB <input type="button" value="v"/>		
Taxonomy	All entries <input type="button" value="v"/>		
Enzyme	Trypsin <input type="button" value="v"/>	Allow up to	1 <input type="button" value="v"/> missed cleavages
Fixed modifications	<input type="text" value="AB_old_ICATd0 (C)"/> <input type="text" value="AB_old_ICATd8 (C)"/> <input type="text" value="Acetyl (K)"/> <input type="text" value="Acetyl (N-term)"/> <input type="text" value="Amide (C-term)"/>	Variable modifications	<input type="text" value="AB_old_ICATd0 (C)"/> <input type="text" value="AB_old_ICATd8 (C)"/> <input type="text" value="Acetyl (K)"/> <input type="text" value="Acetyl (N-term)"/> <input type="text" value="Amide (C-term)"/>
Protein mass	<input type="text"/> kDa	ICAT	<input type="checkbox"/>
Peptide tol. ±	2.0 <input type="text"/> Da <input type="button" value="v"/>	MS/MS tol. ±	0.8 <input type="text"/> Da <input type="button" value="v"/>
Peptide charge	2+ <input type="button" value="v"/>	Monoisotopic	<input checked="" type="radio"/> Average <input type="radio"/>
Data file	<input type="text"/>	<input type="button" value="Browse..."/>	
Data format	Mascot generic <input type="button" value="v"/>	Precursor	<input type="text"/> m/z
Instrument	Default <input type="button" value="v"/>		
Overview	<input type="checkbox"/>	Report top	20 <input type="button" value="v"/> hits
<input type="button" value="Start Search ..."/>		<input type="button" value="Reset Form"/>	

Exercise #2

- Analysis of a human nuclear protein (65 KDa) treated with iodoacetamide and trypsinized followed by MS/MS
- Go to “Worked Example 2” in your notes to follow instructions
- Access your MS/MS data at:
<http://gchelpdesk.ualberta.ca/ABRF2005/>

listed as Example2.dta

Mascot and MS/MS formats

- For MS/MS work, the data file must contain 1 or more sets of MS/MS data
- Supported sets include:
 - * Finnigan (.ASC)
 - * Micromass (.PKL)
 - * Sequest (.DTA)
 - * PerSeptive (.PKS)
 - * Sciex API III
 - * Mascot Generic Format (.MGF)

Mascot Generic Format (MGF)

COM=10 pmol digest of Sample X15

ITOL=1

ITOLU=Da

MODS=Met Ox,Cys B propionamide

MASS=Monoisotopic

USERNAME=Lou Scene

USEREMAIL=leu@altered-state.edu

CHARGE=2+ and 3+

BEGIN IONS

TITLE=Peak 1

PEPMASS=983.6

846.60 73

846.80 44

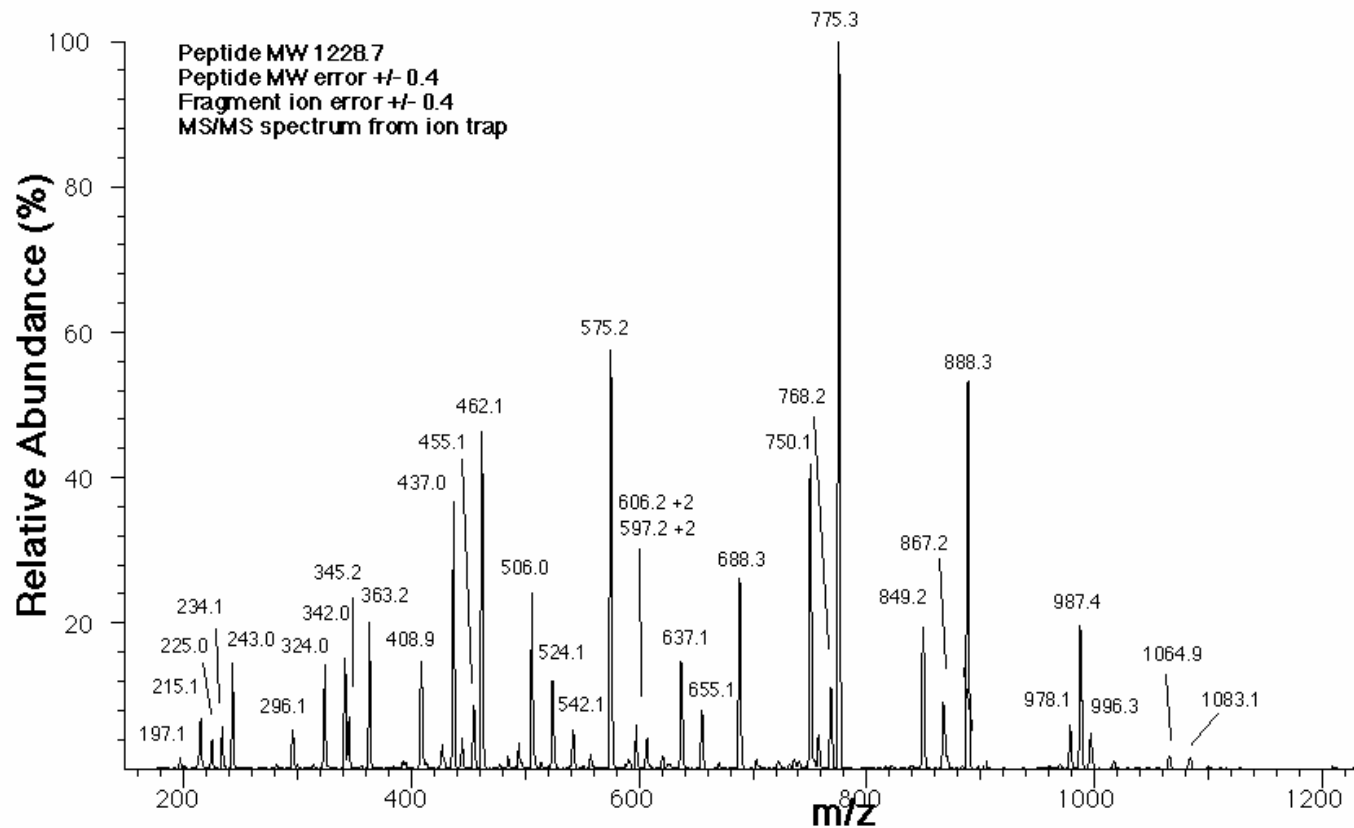
847.60 67

Daughter ion
mass

Parent ion
Mass (2+)

intensity

Example #3 A “Hard” MS/MS Problem

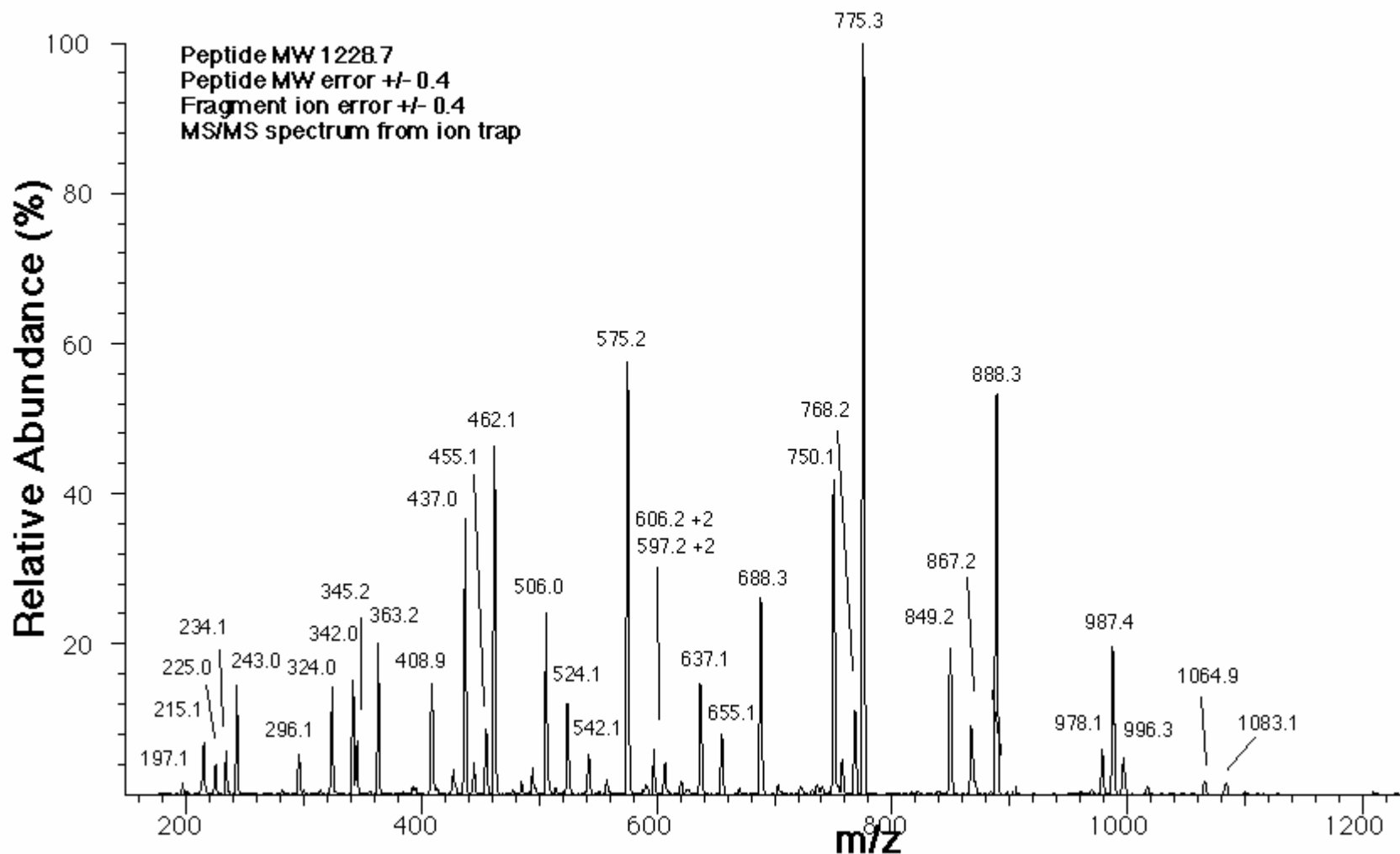


Exercise #3

- Analysis of a novel neuropeptide hormone induced by music/sound
- No known or suspected PTMs
- Ion trap MS-MS spectrum – What is it? What's the sequence?
- Access your MS/MS data at:
<http://gchelpdesk.ualberta.ca/ABRF2005/>

listed as Example3.mgf

MS/MS Spectrum of Neurosensin



What Do You Find?

Protocols for MS-MS Sequencing

- Usually can't tell a "b" ion from a "y" ion
- Assume the lowest mass visible in the spectrum is a lysine or arginine (this is the y_1 ion) this is because trypsin cuts after a lysine or arginine
- This y_1 mass should be 147.113 for lysine or 175.119 for arginine {The y_1 ion is calculated by adding 19.018 u (three hydrogens and one oxygen) to the residue masses of lysine and arginine}

MS-MS Sequencing

- Using the mass tables, look to the right of y_1 and see if you can find another prominent peak that is equal to $y_1 + AA$ where AA is the residue mass for any of the 20 amino acids. This is the y_2 ion
- Proceed in a rightward direction, identifying other y_n ions that differ by an AA residue mass (don't expect to find all)
- The y_n series produces a “reverse” sequence
- Watch for possible dipeptide peaks that may fool you

Things To Remember

- **Gly + Gly = 114.043 u and Asn = 114.043 u**
- **Ala + Gly = 128.059 u and Gln = 128.059 u and Lys = 128.095 u**
- **Gly + Val = 156.090 u and Arg = 156.101 u**
- **Ala + Asp = Glu + Gly = 186.064 and Trp = 186.079 u**
- **Ser + Val = 186.100 u and Trp = 186.079 u**
- **Leu = Ile = 113.084u**

MS-MS Sequencing

- Use the remaining “unassigned” peaks to see if you can construct a “b” ion series
- The highest mass peak corresponds to the parent ion or parent minus 147 (K) or 175 (R)
- The “b” ions give the “normal” sequence
- Both forward (b ion) and backward (y ion) sequences should be consistent
- Use the resulting sequence tag to search the databases using BLAST (remember to use a high Expect value ~ 100) to see if the sequence matches something

Conclusions

- **Mascot is an excellent FREE resource for doing PMF and MS/MS searches of proteins**
- **Understanding the scoring scheme and importance of database size (and mass tolerance) is critical to using Mascot optimally**
- **Not everything can be done on Mascot**