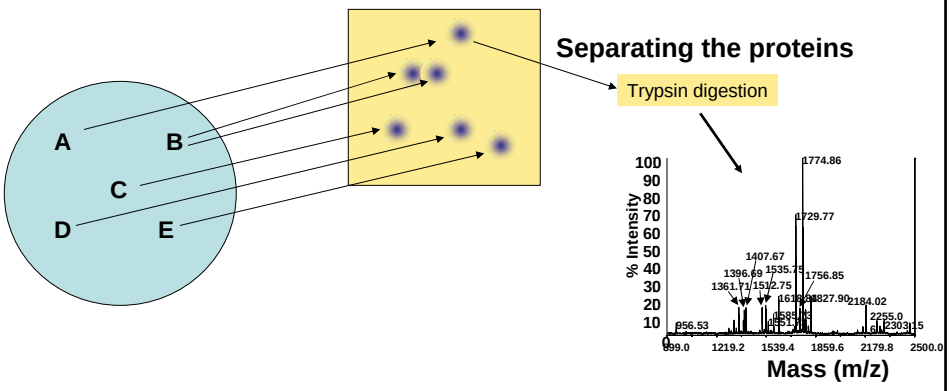
Connecting proteomics into bioinformatics

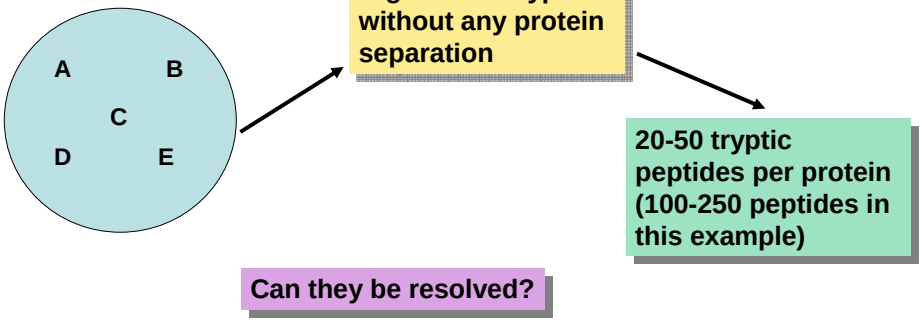
Stephen Barnes
4-7117; sbarnes@uab.edu
Mahyar Sabripour
msabripour@ms.soph.uab.edu

Identifying proteins in a proteome



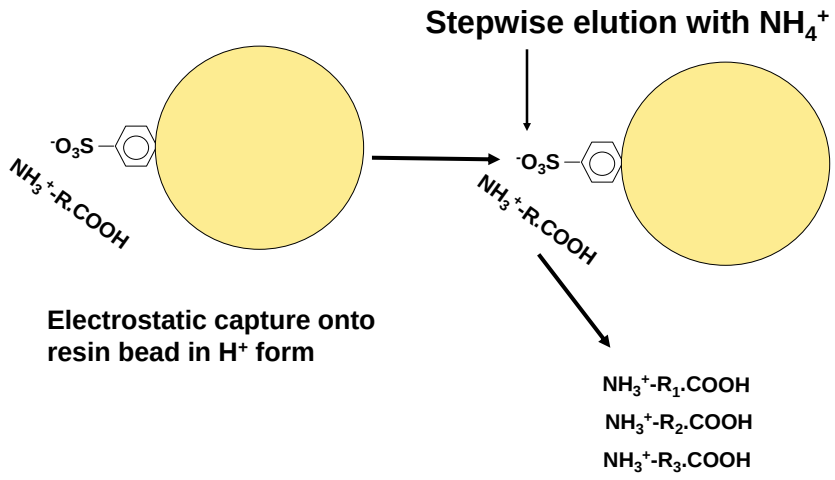
The MUDPIT approach

MUlti-Dimensional Protein Identification Technology



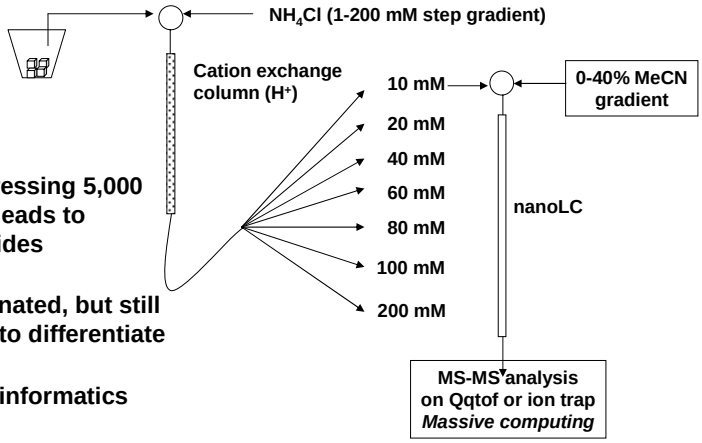
S Barnes/M Sabripour 1/27/06

Cation exchange of peptides



S Barnes/M Sabripour 1/27/06

Fractionation of peptides in MUDPIT analysis



- For a cell expressing 5,000 proteins, this leads to >100,000 peptides
- Can be fractionated, but still 10,000-20,000 to differentiate
- Enormous bioinformatics problem

S Barnes/M Sabripour 1/27/06 **John Yates**

Column construction for MudPIT

100 μm i.d. fused silica capillary

3 cm Polaris C_{18} RP 5 μm

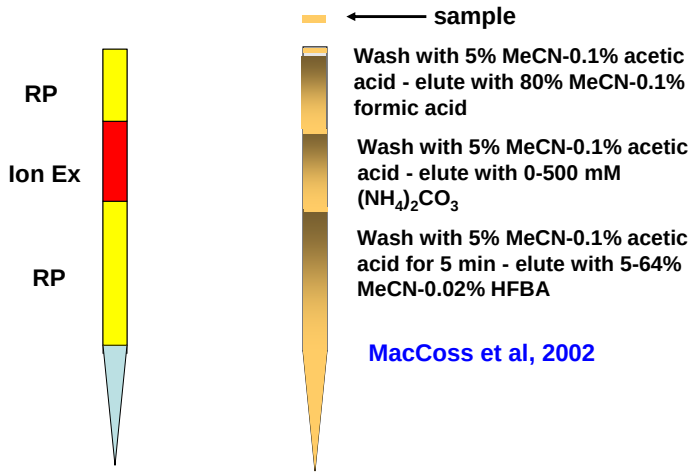
3-6 cm Partisphere SAX 5 μm
Strong cation exchange resin

7 cm Polaris C_{18} RP 5 μm

5 μm tip

S Barnes/M Sabripour 1/27/06 **MacCoss et al, 2002**

Elution from a triphasic column



MacCoss et al, 2002

S Barnes/M Sabripour 1/27/06

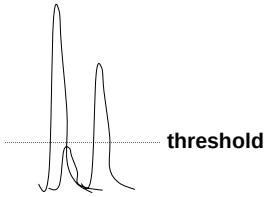
Pros and cons of triphasic columns

- **Pros**
  - **Minimum dead volume between different packings**
  - **Highest sensitivity**
- **Cons**
  - **Have to use NH_4Ac rather than KCl**
  - **Extremely difficult to be reproducible (inconsistent elution times from one column to the next)**

S Barnes/M Sabripour 1/27/06

Issues in MS-MS experiment

- At any one moment, several peptides may be co-eluting
- Data-dependent operation:
  - The most intense peptide molecular ion is selected first (must exceed an initial threshold value)
  - A 2-3 Da window is used (to maximize the signal)
  - The ion must be in 2^+ or 3^+ state
  - Since the ion trap scan of the fragment ions takes ~ 1 sec, only the most intense ions will be measured
  - However, can use an exclusion list on a subsequent run to study minor ions



S Barnes/M Sabripour 1/27/06

The tandem MS mountain

- In a typical MUDPIT experiment, >50,000 tandem MS spectra will be acquired
- Argued that it would take too long to individually interpret each spectrum
  - If it took 15 min per spectrum, then 50,000 spectra would take 12,500 hr, or ~300 wk of effort (6 yr)
- Automated methods were sought that used computer-based comparisons with known databases of protein sequence information
- Development of SEQUEST and MASCOT methods

S Barnes/M Sabripour 1/27/06

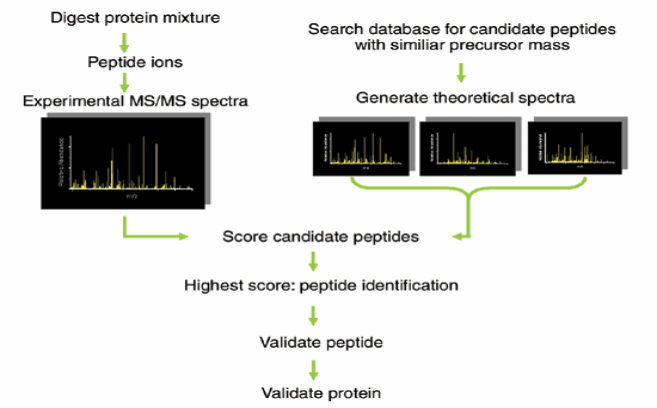
Database Searching and Scoring of Tandem Mass Spectra

- Four main approaches
- Descriptive models (SEQUEST)
- Interpretative models (PeptideSearch)
- Stochastic models (Scope)
- Statistical Models (MASCOT)
- Most commonly used: SEQUEST and MASCOT

Nat Methods. 2004 Dec;1(3):195-202. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Sadygov RG, Cociorva D, Yates JR III.

S Barnes/M Sabripour 1/27/06

Database Searching and Scoring of Tandem Mass Spectra



Sadygov RG et al.

S Barnes/M Sabripour 1/27/06

Scoring tandem mass spectra

- **What does SEQUEST use?**
  - **What is Xcorr?**
- **What is MASCOT?**
  - **Does it have advantages over Xcorr?**

S Barnes/M Sabripour 1/27/06

SEQUEST

- **Descriptive model for comparing MS/MS spectra against observed spectra**
- **Uses peptide mass and cross-correlation score (Xcorr) to rank potential matches**
- **Very computationally intensive**
- **http://fields.scripps.edu/sequest/**

S Barnes/M Sabripour 1/27/06

SEQUEST

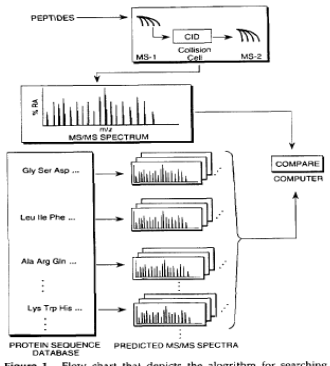


Figure 1. Flow chart that depicts the alogrithm for searching protein databases with tandem mass spectrometry data.

Steps involved:

Step 1: Mass spectrometry data reduction

Step 2: Search method

Step 3: Scoring method

Step 4: Cross-Correlation analysis (Xcorr)

J Am Soc Mass Spectrom. 1994, 5, 976-989. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. Eng, J. K., McCormack, A. L., Yates III, J. R.

S Barnes/M Sabripour 1/27/06

Step 1: Mass Spectrometry Data Reduction

- **Fragment ion *m/z* are converted to nearest integer.**
- **10-u window around precursor ion removed**
- **All but the 200 most abundant ions are removed and remaining ions are renormalized to 100.**

S Barnes/M Sabripour 1/27/06

Step 2: Search Method

- **To match a pair of spectra, protein sequences are retrieved from the database which have masses (within a certain mass tolerance) matching the peptide of interest.**
- ***m/z* values for the predicted fragment ions of each sequence are calculated**

S Barnes/M Sabripour 1/27/06

Step 3: Scoring Method

- Calculate S_p

$$S_p = \left(\sum_k I_k\right) m(1+\beta)(1+\rho)/L$$

where the first term in the product is the sum of ion abundances of all matched peaks, m is the number of matches, β is a 'reward' for each consecutive match of an ion series (for example, 0.075), ρ is a 'reward' for the presence of an immonium ion (for example, 0.15) and L is the number of all theoretical ions of an amino acid sequence.

The higher the value of S_p the better. Larger peptides have larger S_p

S Barnes/M Sabripour 1/27/06

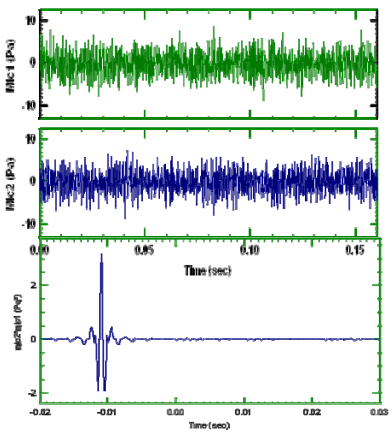
Step 4: Cross-correlation analysis

- **What is cross-correlation?**
- **Method of estimating the degree to which two signals are correlated**
- **Used heavily in time-series analysis and signal processing**
- **Closer XCorr is to 1, the better the match**

S Barnes/M Sabripour 1/27/06

Cross-Correlation Analysis

- Have signal from source 1:
$$X_1(t) = p_1(t) + e_1(t)$$
- Have signal from source 2:
$$X_2(t) = p_2(t) + e_2(t)$$
- **The Cross-correlation (XCorr) shows how similar the two different signals are at various lag times.**



Graphs from www.prosig.com

S Barnes/M Sabripour 1/27/06

Cross-Correlation Analysis

- **SEQUEST reconstructs a mass spectrum from the amino-acid sequences obtained from the database.**
- **The observed spectrum and the reconstructed spectrum are taken and the cross-correlation is obtained via the use of Fourier transforms.**

S Barnes/M Sabripour 1/27/06

Cross-Correlation Analysis

- **Cross-Correlation between two continuous signals**:

$$C_{xy} = \int_{-\infty}^{+\infty} x(t)\, y(t+\tau)\, dt$$

- **For Discrete signals (usually the case)**

$$R_{\tau} = \sum_{i=0}^{n-1} x[i]\, y[i+\tau]$$

- **Usually calculated via FFT (Fast Fourier Transform)**

$$f \star g = \mathcal{F}\,[\overline{F}(\nu)\, G(\nu)].$$

S Barnes/M Sabripour 1/27/06

Sequest Output

#	Rank/Sp	(M+H)+	Cn	deltCn	C*10^4	Sp	Ions	Reference	Peptide
1.	1 / 1	1471.7	1.0000	0.0000	3.8603	851.3	22/39	G3P1_YEAS+4	(R)VPTVDVSVVDLTVK
2.	2 / 8	1469.7	0.6042	0.3958	2.3323	381.5	16/39	S52527	(L)QAPPPPPSSTKSKF
3.	3 / 2	1472.9	0.5877	0.4123	2.2688	448.7	17/39	KEX1_YEAS	(A)VVVTIVTFLIVVLG
4.	4 / 9	1469.6	0.5573	0.4427	2.1515	378.5	17/39	CBS1_YEAS	(R)VPMTGDLSTGNTFE
5.	5 / 12	1471.8	0.5356	0.4644	2.0677	368.2	17/39	ODPA_YEAS	(S)VKAVLAELMGRRAG

normalized correlation score (Cn)
1.0 - normalized correlation score (deltaCn)
tells you how different the first hit is from subsequent hits. (Values > 0.1 indicate a good hit)
raw correlation score (C*10^4).

Normalized XCorr = (XCorr - Cutoff) / Cutoff
Where Cutoff is either 1.8, 2.5, or 3.5 depending on charge state.

S Barnes/M Sabripour 1/27/06

MASCOT

- **Associates a probability with spectrum matches**
- **Implements probability based implementation of the MOWSE scoring algorithm**
- **www.matrixscience.com**

S Barnes/M Sabripour 1/27/06

MOWSE

- **Molecular Weight Search**
- **Scoring is based on peptide frequency distribution within database (frequency factor matrix)**

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

S Barnes/M Sabripour 1/27/06

MOWSE

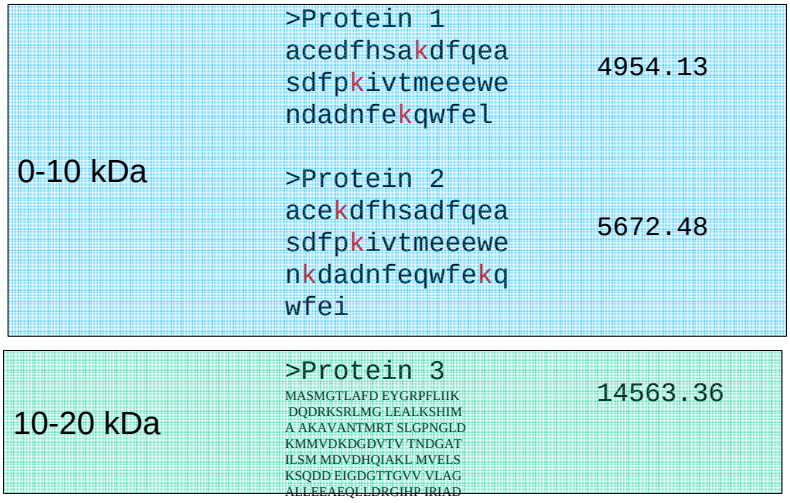
Sequence	Mass (M+H)	Tryptic Fragments
>Protein 1 acedfhsakdfqea sdfpkivtmeeewe ndadnfekqwfe	4842.05	acedfhsak dfgeasdfpk ivtmeeewendadnfek gwfe
>Protein 2 acekdfhsadfqea sdfpkivtmeeewe nkdadnfeqwfe	4842.05	acek dfhsadfgeasdfpk ivtmeeewenk dadnfeqwfe
>Protein 3 MASMGTLAFD EYGRPFLIIK DQDRKSRLMG LEALKSHIM A AKAVANTMRT SLGPNGLD KMMVDKDGDVTV TNDGAT ILSM MDVDHQIAKL MVELS KSQDD EIGDGTTGVV VLAG ALLEEAEQLLDRGIHP IRIAD	14563.36	SQDDEIGDGTTGVVVLAGALLEEAEQLLDR2 DGDVTVTNDGATILSMMDVD HQIAK MASMGTLAFDEYGRPFLIIK2 TSLGPNGLDK LMGLEALK LMVELSK AVANTMR SHIMAAK GIHPIR MMVDK DQDR

From Bioinformatics.ca

S Barnes/M Sabripour 1/27/06

MOWSE

1. Group Proteins into 10 kDa ‘bins’.



From Bioinformatics.ca

S Barnes/M Sabripour 1/27/06

MOWSE

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

>Protein 1
acedfhsakdfqea
sdfpkivtmeeewe
ndadnfekqwfel

>Protein 2
acekdfhsadfqea
sdfpkivtmeeewe
nkdadnfeqwfekq
wfei

Mol. Wt.	Fragment
2098.8909	IVTMEEEWENDADNFEK
1183.5266	DFQEASDFPK
1007.4251	ACEDFHSAK
722.3508	QWFEL
1740.7500	DFHSADFQEASDFPK
1407.6460	IVTMEEEWENK
1456.6127	DADNFEQWFEK
722.3508	QWFEI

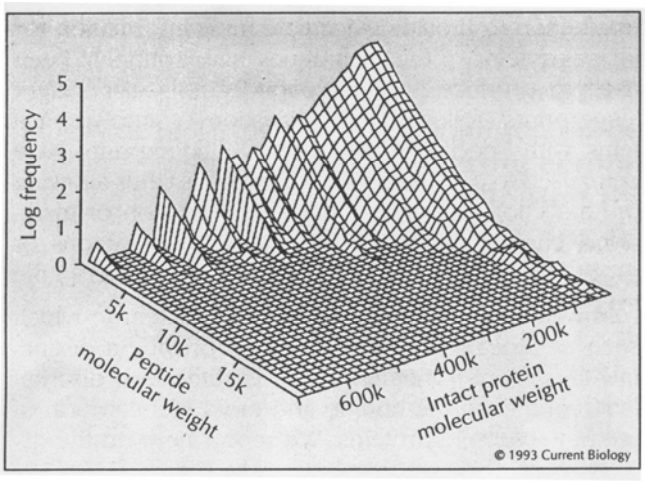
Bin	Fragment
2000-2100	IVTMEEEWENDADNFEK
1900-2000	
1800-1900	
1700-1800	DFHSADFQEASDFPK
1600-1700	
1500-1600	
1400-1500	IVTMEEEWENK, DADNFEQWFE
1300-1400	
1200-1300	
1100-1200	DFQEASDFPK
1000-1100	ACEDFHSAK
900-1000	
800-900	
700-800	
600-700	QWFEL, QWFEI
500-600	
400-500	

From Bioinformatics.ca

S Barnes/M Sabripour 1/27/06

MOWSE

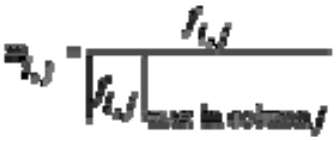
The MOWSE frequency distribution plot looks like this:



From Bioinformatics.ca

MOWSE

- **3. Normalize our frequency factor matrix by the largest value in the column to give the MOWSE factor matrix M**



MOWSE

- 4. **Create MOWSE Score after searching experimental mass values against peptide mass database**.

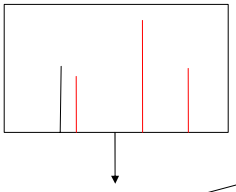
$$Score = \frac{50{,}000}{\sum_{Peak} \prod_{n} I_{n,j}}$$

- **Mprot is the molecular weight of the entry**

S Barnes/M Sabripour 1/27/06

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

Bin	Fragment	Total	Frequency	Normalized
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 x 1 x 1 = 0.5

From Bioinformatics.ca

S Barnes/M Sabripour 1/27/06

MOWSE

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

P_N = product of distribution frequency scores
= 0.5 x 1 x 1 = 0.5

$$\text{Score} = \frac{50\,000}{P_N \times H}$$

H = 'Hit' Protein MW
= 5672.48

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

From Bioinformatics.ca

S Barnes/M Sabripour 1/27/06

MASCOT

- **Details of probability model are not published.**
- **The Mascot Score is given as**

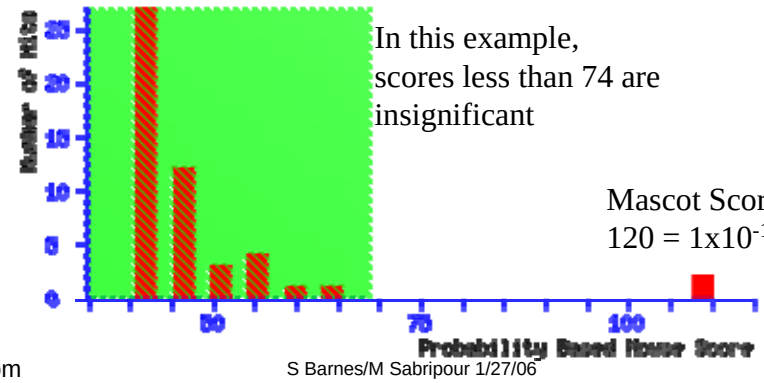
  S = -10*Log(P), where P is the probability that the observed match is a random event

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.

S Barnes/M Sabripour 1/27/06

Mascot Scoring

- The Mascot Score is given as S = -10*Log(P), where P is the probability that the 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.



In this example, scores less than 74 are insignificant

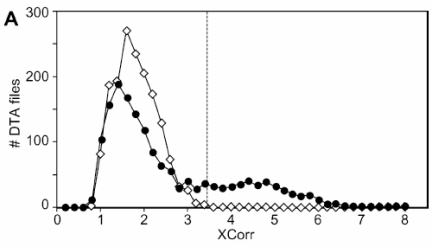
Mascot Score: $120 = 1x10^{-12}$

From Bioinformatics.ca

More haste, less speed?

- Post analysis, the masses of the peptides triggering MS-MS are used to create a set of virtual peptides with masses within ± 1 Da
- Predicted MS-MS are compared to the observed and the best fit is reported as a hit
- The abundance of these hits are plotted in the figure as closed circles

Resing et al. (2004)



However, if the proteome is reversed and sequences of the peptides within ± 1 Da and their predicted MS-MS compared to the observed spectra, a similar histogram is obtained (open circles), but without the right side tail

A forced fit to a set of data will always come up with a match, but not necessarily the truth

Effect of reversing a peptide on the fragment ions that are observed

Normal sequence

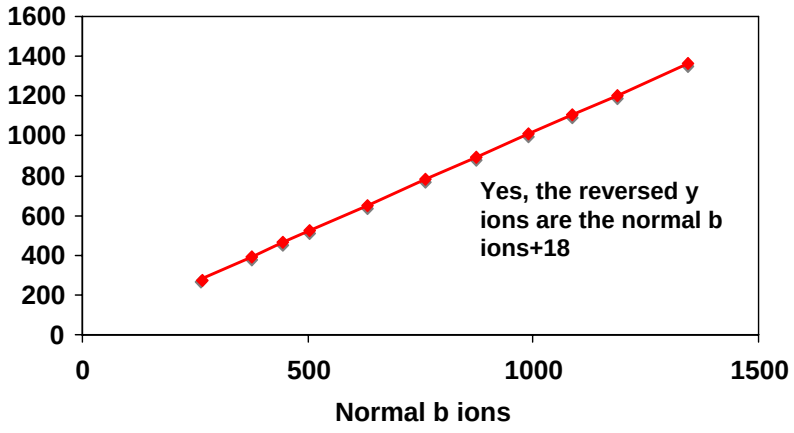
b ions		262	375	446	503	632	760	875	989	1088	1187	1343
	N	F	L	A	G	E	K	D	N	V	V	R
y ions	1361	1247	1100	987	916	859	730	602	487	373	274	175

Reversed sequence

b ions		256	355	469	584	712	841	898	969	1082	1229	1343
	R	V	V	N	D	K	E	G	A	L	F	N
y ions	1361	1205	1106	1007	893	778	650	521	464	393	280	133

S Barnes/M Sabripour 1/27/06

Are the reversed sequence fragment ions correlated with the normal fragment ions?



S Barnes/M Sabripour 1/27/06

Conclusion

- **The reversed peptide sequence generates b ions that are similar to the normal sequence y ions**
  - **Off by 18 Da (i.e., H_2O)**
- **A random peptide sequence library is needed to assess the quality MUDPIT data**

S Barnes/M Sabripour 1/27/06

How to improve MUDPIT

- **Reproducible column engineering**
  - **Tandem columns, each built to separate, but high specifications**
  - **Columns on a chip**
- **More careful selection of the parent ion**
  - **Accurate measurement of the peptide's mass will eliminate many false peptides**
  - **Accurate measurement of peptide fragments' masses**
- **Greater stringency in assessing score cutoff**

S Barnes/M Sabripour 1/27/06