MSMS Database Search

Tim T Chen

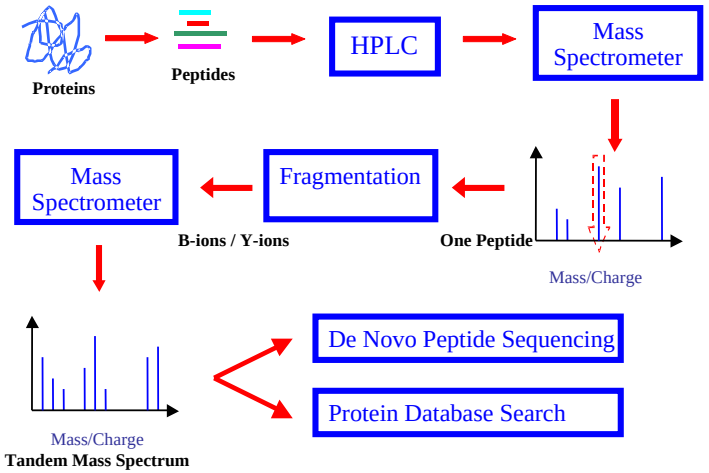
Departments of Biology, Computer Science and Mathematics

University of Southern California

MS based protein identification method

- Finger printing
- MS/MS

Peptide Identification: HPLC-MS-MS



Applications of Mass Spectrometry

- Identification of proteins and protein complexes
- Protein sequencing
- Protein quantitation
- Identification of modified and mutated proteins
- Identification of protein cross-links for protein structure analysis
- Identification of protein-drug interaction
- Selecting mass peaks (proteins) for cancer/disease diagnosis
- …

Common Terminology and Abbreviations:

m/z: mass-to-charge ratio, which is the data reported by the mass spec. When z is known, molecular weight can be determined.

Abundance/Intensity: The number of ions detected**.**

Digestion: Trypson cuts after K and R but not before P.

ESI: Electro-Spray Ionization, a technique for generating charged, gas phase ions from a liquid phase source - great for peptides and used in LC/MS applications.

MALDI: Matrix-Assisted Laser Desorption Ionization, a technique for generate ions from a solid phase source (dried on a plate), good for intact proteins.

Tandem Mass Spec: (aka MS/MS) m/z determination followed by a round of fragmentation and then determination of resulting m/z's. Can be repeated indefinitely - MS^n.

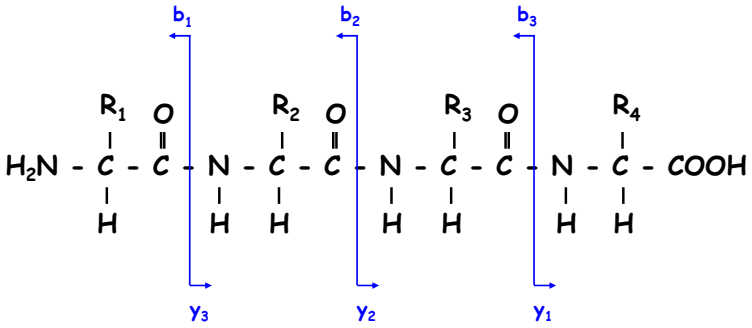
Ion trap: method for electrically retaining an ion of interest while letting all others pass freely out of the mass spec

CID: collision-induced dissociation, a way of causing an ionized peptide to fragment along its peptide bonds

Different Digestion Methods

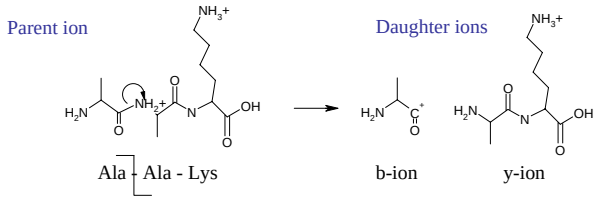
Name	Cleave	Don't cleave	N or C term
Trypsin	KR	P	CTERM
Arg-C	R	P	CTERM
Asp-N	BD		NTERM
Asp-N_ambic	DE		NTERM
Chymotrypsin	FYWLIVM	P	CTERM
CNBr	M		CTERM
Formic_acid	D		CTERM
Lys-C	K	P	CTERM
Lys-C/P	K		CTERM
PepsinA	FL		CTERM
Tryp-CNBr	KRM	P	CTERM
TrypChymo	FYWLKR	P	CTERM
Trypsin/P	KR		CTERM
V8-DE	BDEZ	P	CTERM
V8-E	EZ	P	CTERM
CNBr+Trypsin	M		CTERM
	KR	P	CTERM

MS/MS Sequencing of Peptides



How fragmentation works

•Multiply charged ions are generated at ionization stage
•Protons can migrate along peptide backbone, pausing at peptide bonds
•Excited helium gas collides with charged peptides
•Collision preferentially causes cleavage at peptide bonds, made labile by extra proton



•Helium is excited with voltage tuned to parent ion - one "hit" per peptide

ion masses

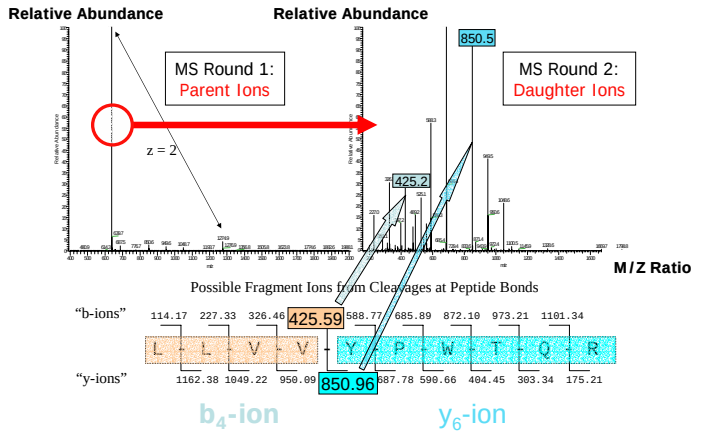
- b: residue mass + 1 proton (1)
- Y: residue mass + 3 protons + 1 oxygen (19)
- Isotopic ions: C12 (99%) and C13 (1%)
- Ion types: b, y, b-H2O, b-NH3, y-H2O, y-NH3, a, x, b-2H2O, y-2H2O, b2+, y2+

	ESI QUAD TOF	MALDI TOF PSD	ESI TRAP	ESI QUAD	ESI FTICR	MALDI TOF TOF	ESI 4 SECT	FTMS ECD	MALDI QUAD TOF
1+ Fragments	×	×	×	×	×	×	×	×	×
2+ Fragments if precursor is 2+ or higher	×		×	×	×		×	×	×
Immonium Ions		×				×	×		×
a series ions		×				×	×		
a-NH3 if fragment includes RKNQ		×				×			
a-H2O if fragment includes STED		×				×			
b series ions	×	×	×	×	×	×	×		×
b-NH3 if fragment includes RKNQ	×	×	×	×	×	×	×		×
b-H2O if fragment includes STED	×	×	×	×	×	×	×		×
y series ions	×	×	×	×	×	×	×	×	×
y-NH3 if fragment includes RKNQ	×		×	×	×	×			×
y-H2O if fragment includes STED	×		×	×	×	×			×
internal yb < 700 Da						×	×		×
internal ya < 700 Da						×	×		×

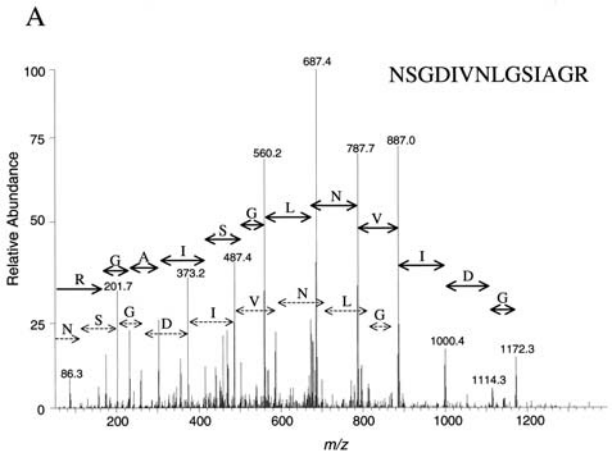
Amino Acid Residue Mass Table (Average)

A	71.08		M	131.19
C	103.14		N	114.1
D	115.09		P	97.12
E	129.12		Q	128.13
F	147.18		R	156.19
G	57.05		S	87.08
H	137.14		T	101.11
I	113.16		V	99.13
K	128.17		W	186.21
L	113.16		Y	163.18

MS/MS – Isolation, Fragmentation, and Determination

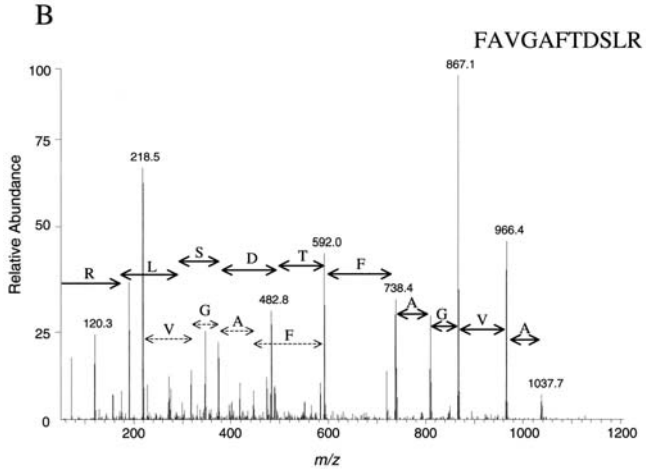


Tandem Mass Spectra Analysis



Gygi *et al*. Mol. Cell Bio. (1999)

Tandem Mass Spectra Analysis



Gygi *et al*. Mol. Cell Bio. (1999)

Mass Spectrometry - What it can do:

- Determine m/z - mass to charge ratio (& ∴ molecular weight) very accurately
- Verify proper peptide synthesis/expression product
- Identify unknown protein from gel bands
- Identify components of protein mixtures

} Greatly aided by presence in a sequence database

- De novo sequencing of peptides
- Quantitation of product by comparison (ICAT) or when internal standard is present

Mass Spectrometry - What it can't do:

- Differentiate compounds with identical molecular weights (e.g. Leu vs. Ile in peptide sequencing)
- No guarantees - analysis dependent on ability to ionize analyte multiple ionization needed for sequence analysis
- Identify every component in a gmish in one shot
- Interpret your data for you

Mass Spectrum Interpretation Challenge

- It is unknown whether an ion is a b-ion or an y-ion or else.
- Some ions are missing.
- Each ion has a couple of isotopic forms.
- Other ions (a or z) may appear.
- Some ions may lose a water or an ammonia.
- Noise.
- Amino acid modifications.

Database Searching Using MS/MS data

- Input: a MS/MS spectrum and a protein sequence database;
- Output: The peptide in the database that can explain the MS/MS spectrum

Protein Identification Problem (PID):

Given a database D, a mass W, an error range e, and a spectrum S, ask for a sequence P from D such that
(1)| mass(P)—W| < e, and
(2) Let T be a set of all ion masses (prefix/suffix sums) of P.
Then S and T are optimally correlated.

Given
e = 0.5
W = 429.100
S = {199.022,274.31,361.01}
D = {MCAKSWRYIL…}

→

$\boldsymbol{P}$ = SWR,
Mass($\boldsymbol{P}$) = 429.212,
B-ions($\boldsymbol{P}$) ={88.033,274.112}
Y-ions($\boldsymbol{P}$)={175.113,361.121}
T ={88.033,175.113,274.112,361.121}

Step 1. Preprocess the protein database

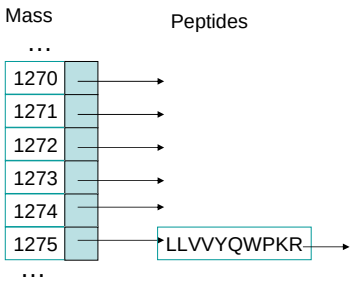
- If the enzyme is known, the protein database can by preprocessed by digestion and indexing.
- Example:

  If the enzyme is “trypsin”, then the protein sequences can be digested by the computer according to the rule:

  after R (Arginine) and K (Lysine),
  but not before P (Proline).

  The digested peptides can then be indexed by the mass..

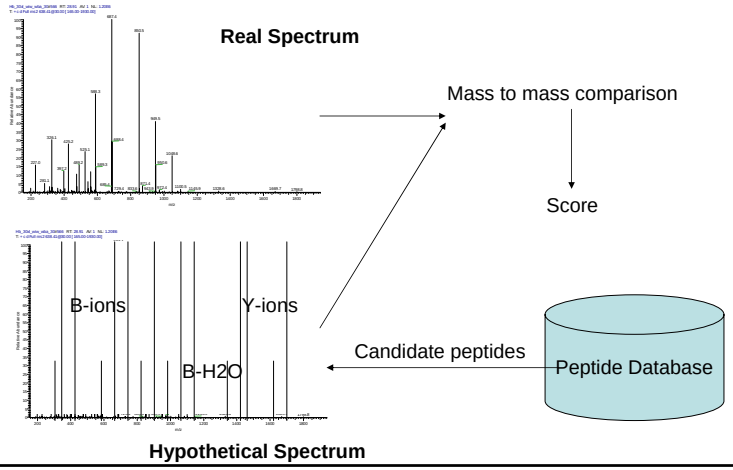
Mass Array

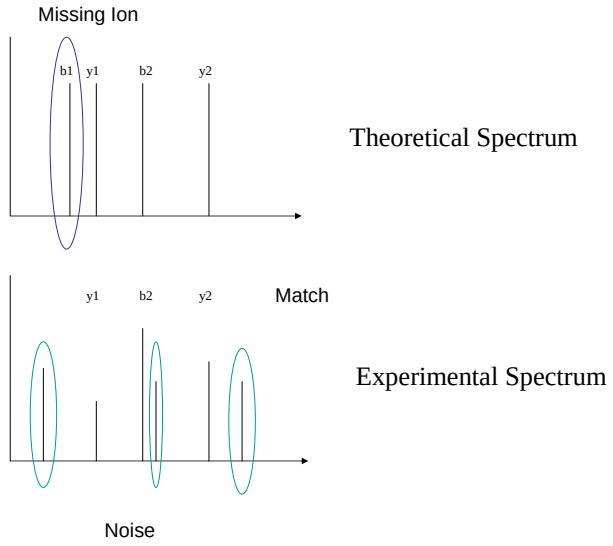


Step 2 Search using parent ion mass

- Search the indexed peptide database using the parent ion mass of the query tandem mass spectrum.

Step 3 Spectrum Comparison





Step 4 Scoring Function

- Sequest: correlation score
- Mascot: probability score
- SCOPE
- Decision Tree and Bayesian Net

SEQUEST Scoring Method

Scoring each peptide by comparing the hypothetical spectrum with the experimental spectrum

- One simple score:

$$S_p = (\sum_{m:Matching} i_m^H i_m^R) n_m (1+b)(1+r) / n_R$$

- Cross correlation score

$$R_t = \sum_{i=1}^{n} x[i] y[i+t]$$

From *Yates et al.* Analytic Chemistry 1995

Probability-based Mascot Scoring Method

Calculate the probability that the observed match between the experimental data and each sequence database entry is a chance event. Report a score which is -10Log(*p*), where *p* is the probability.

From *Perkins et al,* Electrophoresis 1999

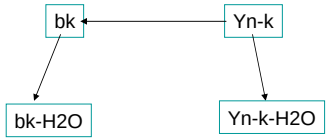
SCOPE

$$\max_p \Pr(S \mid F) \Pr(F \mid p)$$

p: peptides
F: fragmented ions
S: spectrum

Bafna and Edwards *Bioinformatics* 2002.

Decision Tree and Bayesian Net



Elias et al. Nat Biotechnology 2004

Difficulties

- Unknown fragmentation patterns
- Different kinds of ion series which are machine dependent.
- Different enzyme digestion methods
- Unknown Modifications
- Underestimated mass measurement error
- Incorrect determination of precursor charge
- Peptide sequence not in the database
- Separate signals from noises

Reference: scoring function

1. Eng et al. An Approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *American Society for Mass Spectrometry*, 5:976-989, 1994.
2. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999 Dec;20(18):3551-67.
3. Bafna and Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics. 2001;17 Suppl 1:S13-21.
4. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol. 2004 Feb;22(2):214-9. Epub 2004 Jan 18.

Reference: applications

1. Aebersold et al. Mass Spectrometry in Proteomics. *Chem. Rev.,* 101:269-295, 2001.
2. Gygi, S.P. et al. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *PNAS*, 97(17):9390-9395.
3. Gygi, S.P. et al. 1999. Correlation between Protein and mRNA Abundance in Yeast. *Molecular and Cellular Biology*, 19(3):1720-1730.
4. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 1999 Oct;17(10)994-999.
5. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR 3rd. Direct analysis of protein complexes using massspectrometry. *Nature Biotechnology*. 1999 Jul;17(7):676-82.
6. Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, Dollinger G. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *PNAS*, 97(11):5802-6, 2000.
7. Ho Y, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. 2002 Jan 10;415(6868):180-3.
8. Gavin AC, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002 10;415(6868):141-7.