Interpretation of MS-Based Proteomics Data

Yet-Ran Chen, 陳逸然

Agricultural Biotechnology Research Center
Academia Sinica

Brief Overview of Protein Identification Workflow

1. Protein Sample
2. Specific Protein Fragmentation
3. Chemical Enzymatic High Energy CID
4. Peptides
5. MS Analysis
6. MS Spectrum
7. MS/MS Analysis of Peptide Fragment
8. Calculation of Peptide Mass
9. Calculation of Peptide Fragment Mass
10. Protein Identification

MS Spectrum $\rightarrow$ ? Mass
Basics for the Mass Spectrometric Data
Basics for the Mass Spectrometric Data

Three Major Definitions of Mass

- Monoisotopic mass: mass of an ion for a given empirical formula calculated using the exact mass of the most abundant isotope of each element

- Average mass: mass of an ion for a given empirical formula calculated using the average atomic weight averaged over all the isotopes for each element

- Nominal mass: mass of an ion with a given empirical formula calculated using the integer mass of the most abundant isotope of each element

The Isotopic Pattern

<table>
<thead>
<tr>
<th>m/z</th>
<th>Relative Abundance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>569.3997</td>
<td>100</td>
</tr>
<tr>
<td>570.3997</td>
<td>27</td>
</tr>
<tr>
<td>571.2997</td>
<td>6</td>
</tr>
</tbody>
</table>

Nominal: 569
Monoisotopic: 569.39997
Average: 569.72386
Most abundant: 569.39997
Basics for the Mass Spectrometric Data

Definition of MS Resolution

$$R = \frac{m}{\Delta m} = \frac{m}{\text{FWHM}}$$

FWHM (Full Width at Half Maximum)

Basics for the Mass Spectrometric Data

Molecular Weight Determination – Low Resolution vs. High Resolution
Basics for the Mass Spectrometric Data

Molecular Weight Determination – Single and Multiple Charged MS Spectrum

**Single Charged**

\[
P_1 = \text{Monoisotopic Mass} + H \\
\Delta (m/z) = 1
\]

**Double Charged**

\[
P_1 = \text{Monoisotopic Mass} + 2H \\
\Delta (m/z) = 0.5
\]

Charge State \( Z = \frac{1}{p_2-p_1} \) then Molecular Weight \( MW = p_1 * Z - Z = (p_1-1)*Z \)

Basics for the Mass Spectrometric Data

Molecular Weight Determination – Unresolved Isotopic Peaks

**Unresolved Isotopic Peaks**

\[
190.1 \\
379.2
\]
Basics for the Mass Spectrometric Data

Choosing Valid Spectra Peaks for the Protein ID

Resolving Power

Ionization Method

Protein Identification by Database Searching
Protein Identification by Database Searching

How to Produce Confidence Experimental Result

- Use common search algorithm(s) and database to search
- Employ method(s) for evaluation of the FP rate
- Plan to integrate data for searching to find weak associations not evident in single datasets
- Employ common/consistent annotation of results
- Store data in original instrument vendor format in as minimally processed form as possible (exchange formats in flux)
  - Files contain all the interesting info in unprocessed form
    - parent peak intensities for quantitation
    - resolution, peak spacing (charge states)
    - acquisition parameters

Etc........

Guideline Draft for Proteomics Data Publication
Molecular Cellular Proteomics, July 14, 2005
The Peptide Mass Fingerprinting (PMF)

The Peptide Mass Fingerprinting

Figure 1

Protein identification using peptide mapping information. (a) In the experiment, the proteins are digested with an enzyme and the masses of the proteolytic peptides are measured with mass spectrometry. (b) In the database search, each protein sequence in the database is digested according to the specificity of the enzyme. The masses of the resulting peptides are calculated and a theoretical mass spectrum is constructed. The measured mass spectrum is compared with the theoretical mass spectrum.
Data Flow of Peptide Mass Fingerprinting

- Digesting the complete database with the enzyme in question.
- Calculating the peptide masses.
- Finding the location with most fitting masses.
- Problems
  - Some masses are quite abundant
  - Not annotated genomes could pose problem

Mass Distribution of Tryptic Peptide

The genomic databases were translated in all six reading frames on the fly and then digested with Trypsin (R|K).
The MOWSE Algorithm

- Relative abundance of peptides of a given length produced by proteolysis depends upon lengths of both peptide and protein.

- In an infinitely long protein the fractional abundance of trypsin peptides of N residues is \(A(1-A)^{N-1}\). A is the abundance of cleaveable bonds.

D.N. Perkins et al., Electrophoresis 1999, 20, 3551-3567

The MOWSE Algorithm

“real” proteins display an “end effect” that increases the abundance of short peptides and the probability of detecting a peptide that has the same length as the protein (so no cleavage has occurred).

D.N. Perkins et al., Electrophoresis 1999, 20, 3551-3567
The MOWSE – Scoring Scheme I

Protein MW

ProteinDB

In Silico Digest

Peptides Generated from 90-100kDa of ProteinDB

Peptide MW

2.9kDa

3kDa

The MOWSE – Scoring Scheme II

Cell Freq. = # of peptides in specific MW / # of total peptides

F-max = most freq. peptide #

F - Cell Frequency
m - Cell Value
Mp - 'hit' protein molecular weight
Average Protein MW ~ 50000 Da
### Evaluation of the Search Result Quality

Simulations provide a method for determining the quality of the search results [41**].
Testing the Significant in Protein Assignment


Distribution of Identification Score

Random Hit

Significant Hit?
The MASCOT Search Engine

Probability-Based implementation of the MOWSE algorithm

The total score is the absolute probability (P) that the observed score is a random event.

The Mascot score is given as -10xlog(P)

MASCOT PMF Search Result

User: A. S. M. S. poster
Email: example@matrixscience.com
Search title: Mass tolerance example
Database: NRDB 30020121 (80000 entries; 255696937 residues)
Timestamp: 24 Feb 2002 at 16:54:14 GMT
Top Score: 109 for HS26_YEAST, Heat shock protein 26 (26 kDa heat shock protein)

Probability Based Mowse Score

The score is -10xlog(P), where P is the probability that the observed match is a random event.
Protein scores greater than 72 are significant (p<0.05).

Search Parameters

Type of search: Peptide Mass Fingerprint
Enzyme: Trypsin
Mass values: Monoisotopic
Protein Mass: Unrestricted
Peptide Mass Tolerance: ± 0.1 Da
Peptide Charge State: 1-
Max Missed Cleavages: 1
Number of spectra: 18
MASCOT PMF Search Result

User: A. S. M. S. poster
Email: example@matrixscience.com
Search title: Mass tolerance example
Database: MSDB 200920121 (320027 sequences: 255696937 residues)
Timestamp: 24 Feb 2002 at 16:34:14 GMT
Top Score: 109 for H526 YEAST, Heat shock protein 26 (26 kDa heat shock protein)

Probability Based Mouse Score

Probability-Based

The Mascot score is given as -10xlog(P)
P: probability that the observed score is a random event

Significant Threshold

Search Parameters

Type of search: Peptide Mass Fingerprint
Enzyme: Trypsin
Mass values: Monoisotopic
Protein Mass: Unrestricted
Peptide Mass Tolerance: ±0.1 Da
Peptide Charge State: ±1
Max Missed Cleavages: 1
Number of queries: 18

Interpretation PMF Search Result

1. AJLHIA Mass: 32244 Score: 167 Expect: 2.1e-012 Queries matched: 11
serum amyloid P component, chain A - human

VI7277 Mass: 26971 Score: 163 Expect: 6.4e-012 Queries matched: 11
serum amyloid P-component precursor (validated) - human

AAAC1011 Mass: 22591 Score: 159 Expect: 1.0e-012 Queries matched: 10
HNS135A BID - Homo sapiens

NAM1734 Mass: 160816 Score: 65 Expect: 4 Queries matched: 9
HMG19C100 NCO - Homo sapiens

2. SCFF1241 Mass: 12044 Score: 55 Expect: 22 Queries matched: 5
C03534 Human

CGFM005140033 Mass: 100670 Score: 30 Expect: 22 Queries matched: 4
Plasminogen heavy chain variable region (fragment) - Homo sapiens (Human)

AJ621174 Mass: 13817 Score: 35 Expect: 22 Queries matched: 4
Lactoferrin N-terminal half-molecule mutants D12L - human

3. AF0001405 Mass: 3603 Score: 35 Expect: 4 Queries matched: 3
T-CELL RECEPTOR DELTA (CHAIN) (FRAGMENT) - Homo sapiens (Human)

Abd11443 Mass: 10255 Score: 35 Expect: 4 Queries matched: 3

Loss scores -10xLog(P), where P is the probability that the observed score is a random event. Scores greater than 72 are significant (p<0.05).
What p-value is significant?

- The most common thresholds are 0.01 ~ 0.05.
- A threshold of 0.05 means you are 95% sure that the result is significant.
- Is 95% enough? It depends upon the cost associated with making a mistake.
- Examples of costs:
  - Doing expensive wet lab validation.
  - Making clinical treatment decisions.
  - Misleading the scientific community.
The Peptide Mass Fingerprinting

- Fast, simple analysis
- High sensitivity
- Protein sequence database required
- Not suitable for mixtures or low MW proteins

Peptide Identification from the MS/MS Spectra
### Issues to be Considered

Apportionment of Proteins from ID peptides
Evaluation of false positive rate of ID peptide
Degenerate peptides are more prevalent with databases of higher eukaryotes due to the presence of:
- related protein family members
- alternative splice forms
- partial sequences
- DB redundancy
- Single hit

### Peak Extraction

![Peak Extraction Diagram]
How we extract the m/z info from raw spectra??

Matching Accuracy Using Different Centroid Settings

15 % Height

50 % Height

95 % Height

Calibration without centroid
(Calibrate by the Peak Apex)

Calibration with centroid
(Centroid Peak Height: 50%)

(Can we use raw data??)
Selection of Search Algorithm

MASCOT

*Probability Based MOWSE Scoring*

- The homology threshold is an empirical measure of whether the match is an outlier

![Graph showing the distribution of scores with a peak at a certain score.]

**ASMS 2005**

Selection of Search Algorithm

SEQUEST

- generate a predicted spectrum for each potential peptide using a simple fragmentation model (all b and y ions have the same intensity; possible losses from b and y have a lower intensity)

![Graph showing the distribution of m/z values with peaks at different intensities.]

- compute a "cross-correlation" score and find the best-matching peptide
- since this operation is very time-consuming, a simpler preliminary score is used to find the 500 peptides in the database that are most likely to be the correct identification
Selection of Search Algorithm

SEQUEST

Step 1: Mass spectrometry data reduction

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

Step 2: Peptide Mass Matching

- 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

Step 3: Preliminary Scoring

\[ S_p = (\sum l_k) m (1+\beta)(1+p)/L \]

- \( m \): number of matches (within \( \pm 1u \))
- \( \beta \): reward consecutive matched ion series (for example, 0.075)
- \( \rho \): reward of immonium ion (for example, 0.15)
- \( L \): is the number of all theoretical ions of an amino acid sequence.

Step 4: Cross-Correlation Analysis (Xcorr)

\[ R_\tau(E,T) = \sum X_i Y_{i+\tau} \quad \tau \rightarrow Xcorr \]

Xcorr of a candidate Value of \( R_\tau \) when \( \tau = 0 \) minus the mean of \( R_\tau \) (over the range -75<\( \tau <75 \) and divided by 10^4)

Selection of Search Algorithm

SEQUEST - The \( R_\tau \)

- \( b \) and \( y \) ions: 50
- \( b \) and \( y \) ions \( \pm 1 \) u (isotopic): 25
- Ions with neutral loss of H₂O, NH₃, CO and \( \pm 1 \) u: 10
Selection of Search Algorithm

\[ R_{\tau}(E,T) = \sum X_i Y_{i+\tau} \]

Time Consuming: For a P4 3.2GHz PC, 1 spectrum takes 1-2 sec, 10,000 spectra (2.7-5.4 hr)

Mascot v.s. SEQUEST

Proteomics 2004, 4, 619–628
Each search engine identifies about the same number of spectra, but the overlap is surprisingly small. Different search engines match different spectra.

The reason that they identify different spectra is because each program has different strengths.

Selection of Search Algorithm

Brian C. Searle, Proteome Software Inc
Selection of interrogate protein/genomic DB

<table>
<thead>
<tr>
<th>Database</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>EST</td>
<td>EST divisions of Genbank, (currently EST_human, EST_mouse, EST_others)</td>
</tr>
<tr>
<td>MSDB</td>
<td>Comprehensive, non-identical protein database</td>
</tr>
<tr>
<td>NCBInr</td>
<td>Comprehensive, non-identical protein database</td>
</tr>
<tr>
<td>OWL</td>
<td>Non-identical protein database (obsolete)</td>
</tr>
<tr>
<td>Random</td>
<td>Random sequences for verifying scoring statistics</td>
</tr>
<tr>
<td>SwissProt</td>
<td>High quality, curated protein database</td>
</tr>
</tbody>
</table>

DB Cross-reference
DB Redundancy
DB Accuracy
DB Size
## Instrument Type → Fragmentation Pattern

<table>
<thead>
<tr>
<th>Instrument Type</th>
<th>EST</th>
<th>ESI Quad</th>
<th>TOF</th>
<th>MALDI-TOF PSA</th>
<th>FTMS</th>
<th>MALDI-TOF SEQ</th>
<th>MALDI-TOF QToF</th>
<th>PROGNOSTIC</th>
<th>TOF</th>
<th>MALDI-TOF PSD</th>
<th>FTMS</th>
<th>MALDI-TOF SEQ</th>
<th>MALDI-TOF QToF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1° fragments</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2° fragments</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>a series ions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>b series ions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>c series ions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>d or d series ions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>e or w series ions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Internal y &lt; 700 Da</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Internal y &lt; 700 Da</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>X or y must be significant</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>zH series ions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>z series ions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>y or y* must be top scoring series</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

### Interpretation of Search Result

**SEQUEST:**
- Xcorr $> 2.0$
- $\Delta C_n > 0.1$

**MASCOT:**
- Score $> 45$

**Threshold?**

Use of conservative scoring and filtering thresholds reduces number of miss-assigned peptides and proteins, but does not eliminate false positives.
Probability Accuracy with Different Algorithms

Amplification of False Positive Error Rate from Peptide to Protein Level

Peptide Level: 50% False Positives

Protein Level: 71% False Positives

Alexey Nesvizhskii, Institute of Systems Biology
Score Distribution → Small and Large Scale

1~10 proteins

>100 proteins

ASMS Workshop and User Meeting 2005

Large Scale Proteomics Data Analysis
Difficulty Interpreting Protein Identification

- Different search score thresholds used to filter data
- Unknown and variable false positive error rates
- No reliable measures of confidence

Testing the Significance in Protein Assignment

Search Against Normal and Decoy DB

Normal Swissprot DB

Decoy Swissprot DB

1. Search

   - Prepare decoy database
   - Search against decoy database
   - Analyze results

The Decoy DB

Decoy DB (Randomized or Reversed)

- Randomized
  - PMF Applicable
  - Peptide number changed after randomized (alter the trial number)
  - Degree of Random

- Reversed
  - Not suitable for PMF
  - Not suitable for no enzyme restricted

The Decoy DB

Decoy DB (Randomized or Reversed)

- Randomized
  - PMF Applicable
  - Peptide number changed after randomized (alter the trial number)
  - Degree of Random

- Reversed
  - Not suitable for PMF
  - Not suitable for no enzyme restricted

Normal sequence

Reverse sequence
Search Against Normal + Decoy DB

Large Scale Proteomics Data Analysis

- incorrect hit ---
- correct hit---

Normal Distribution/Poisson Fitting
  - Mascot
  - X!Tandem

number of spectra

database search score
**ID Score Distribution of Identified Peptides**

entire dataset:

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Peptide</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum 1</td>
<td>LGEYGH</td>
<td>4.5</td>
</tr>
<tr>
<td>Spectrum 2</td>
<td>FQSEEQ</td>
<td>3.4</td>
</tr>
<tr>
<td>Spectrum 3</td>
<td>FLYQE</td>
<td>1.3</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Spectrum N</td>
<td>EIQQKF</td>
<td>2.2</td>
</tr>
</tbody>
</table>

MS/MS spectrum

best match

database search score

---

**Statistical Model of Peptide Prophet**

entire dataset:

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Peptide</th>
<th>Score</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum 1</td>
<td>LGEYGH</td>
<td>4.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Spectrum 2</td>
<td>FQSEEQ</td>
<td>3.4</td>
<td>0.97</td>
</tr>
<tr>
<td>Spectrum 3</td>
<td>FLYQE</td>
<td>1.3</td>
<td>0.01</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Spectrum N</td>
<td>EIQQKF</td>
<td>2.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

probability

unsupervised learning

EM mixture model algorithm learns the most likely distributions among correct and incorrect peptide assignments given the observed data

Discriminating Power of Peptide Prophet

**Sensitivity**: fraction of all correct results passing filter

**Error Rate**: fraction of all results passing filter that are incorrect

Improved discrimination:
more identifications (for the same error rate)


---

MS-Based Proteomics: 2D v.s. shotgun

2D-gel based

protein sample

select and isolate spots

identify proteins by mass spectrometry

shotgun

separate peptide mixture

identify peptides (proteins) by tandem mass spectrometry

Alexey Nesvizhskii, Institute of Systems Biology
Analysis of Complex Protein Mixture

Protein Assignment from Identified Peptide(s)

Correspond to more than a single entry in protein database

protein A or protein B ??
Or both?

In shotgun proteomics the connectivity between peptides and proteins is lost

Degenerate peptides are more prevalent with databases of higher eukaryotes due to the presence of:
- related protein family members
- alternative splice forms
- partial sequences
Protein Assignment from Identified Peptide(s)

Protein Assignment from Identified Peptide(s)
Evaluation of Protein Probability

Apportionment of Degenerate Peptides using Peptide Prophet

** Initialize:**

- \( P_A = P_B \)
- \( \text{wt}_i^A = \text{wt}_i^B = 0.5 \)
- \( \text{wt}_2^B = 1 \)

** Run EM:**

- \( P_2 = 1 - \prod_i (1 - \text{wt}_i^B (1 + \text{D}_i)) \)

- \( \text{wt}_1^A = 0 \)
- \( P_A = 0 \)

- \( \text{wt}_1^A = 1 \)
- \( P_B = 1 - (1 - 0.8)(1 - 0.7) = 0.94 \)

Protein B is present in the sample

Alexey Nesvizhskii, Institute of Systems Biology
### Example for the Protein Apportionment

**ProteinProphet output**

<table>
<thead>
<tr>
<th>#</th>
<th>Protein</th>
<th>P</th>
<th>peptide 1</th>
<th>peptide 2</th>
<th>peptide 3</th>
<th>peptide 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Prot A</td>
<td>P=1</td>
<td>*wt=1</td>
<td>*wt=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prot B</td>
<td>P=1</td>
<td>*wt=1</td>
<td></td>
<td>*wt=1</td>
<td></td>
</tr>
</tbody>
</table>

*: no other sequence database entry has this peptide (wt=1)

**distinct proteins**

<table>
<thead>
<tr>
<th>proteins</th>
<th>peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>B</td>
<td>- - -</td>
</tr>
</tbody>
</table>

**Example for the Protein Apportionment**

**ProteinProphet output**

<table>
<thead>
<tr>
<th>#</th>
<th>Protein</th>
<th>P</th>
<th>peptide 1</th>
<th>peptide 2</th>
<th>peptide 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Prot A</td>
<td>P=1</td>
<td>*wt=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prot B</td>
<td></td>
<td></td>
<td>*wt=1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*wt=1</td>
</tr>
</tbody>
</table>

*: no other sequence database entry has this peptide (wt=1)

**equivalent proteins**

<table>
<thead>
<tr>
<th>proteins</th>
<th>peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1 2 3</td>
</tr>
<tr>
<td>B</td>
<td>- - -</td>
</tr>
</tbody>
</table>
### Example for the Protein Apportionment

**peptides**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**B is a subset protein**

Occam’s razor: Prot A is present in the sample. No conclusive evidence for the presence of Prot B.

**ProteinProphet output**

<table>
<thead>
<tr>
<th>#</th>
<th>Prot</th>
<th>P</th>
<th>wt</th>
<th>peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Prot A</td>
<td>1</td>
<td>1</td>
<td>peptide 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>peptide 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>peptide 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>(end of the list)</strong></td>
<td></td>
</tr>
<tr>
<td>#2</td>
<td>Prot B</td>
<td>0</td>
<td>0</td>
<td>peptide 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>peptide 3</td>
</tr>
</tbody>
</table>

### Example for the Protein Apportionment

**peptides**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**differentiable proteins**

**ProteinProphet output**

<table>
<thead>
<tr>
<th>#</th>
<th>Prot</th>
<th>P</th>
<th>wt</th>
<th>peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Prot A</td>
<td>1</td>
<td>1</td>
<td>peptide 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>peptide 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>peptide 3</td>
</tr>
<tr>
<td>#2</td>
<td>Prot B</td>
<td>1</td>
<td>0.5</td>
<td>peptide 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>peptide 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>(end of the list)</strong></td>
<td></td>
</tr>
</tbody>
</table>
### Example for the Protein Apportionment

<table>
<thead>
<tr>
<th>Peptides</th>
<th>ProteinProphet Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4</td>
<td></td>
</tr>
</tbody>
</table>

**Proteins**
- A
- B
- C

A, C: differentiable proteins
B: subsumable protein

### Example for the Protein Apportionment

<table>
<thead>
<tr>
<th>Peptides</th>
<th>ProteinProphet Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3</td>
<td></td>
</tr>
</tbody>
</table>

**Proteins**
- A
- B
- C

**Protein group**
A has all three peptides (most likely protein). However, B and C combined can account for all the observed peptides as well

<table>
<thead>
<tr>
<th>Protein group</th>
<th>ProteinProphet Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a</strong></td>
<td></td>
</tr>
<tr>
<td>Prot A</td>
<td>Protein group P=1</td>
</tr>
<tr>
<td>wt=1</td>
<td>peptide 1</td>
</tr>
<tr>
<td>wt=1</td>
<td>peptide 2</td>
</tr>
<tr>
<td>wt=1</td>
<td>peptide 3</td>
</tr>
<tr>
<td><strong>b</strong></td>
<td>Prot B P=0</td>
</tr>
<tr>
<td>wt=0</td>
<td>peptide 1</td>
</tr>
<tr>
<td>wt=0</td>
<td>peptide 3</td>
</tr>
<tr>
<td><strong>c</strong></td>
<td>Prot C P=0</td>
</tr>
<tr>
<td>wt=0</td>
<td>peptide 2</td>
</tr>
<tr>
<td>wt=0</td>
<td>peptide 3</td>
</tr>
</tbody>
</table>
Output of ProteinProphet

Protein Apportionment with MASCOT
Protein Apportionment with MASCOT

Removing Degenerated Peptides from MASCOT
Search Result
MS/MS Database Search

Easily automated for high throughput
More Accurate then PMF
Can get matches from marginal data
Time consuming
  - No enzyme restriction
  - Many variable modifications
  - Large database
  - Large dataset
MS/MS is peptide identification
  - Proteins by inference
Protein sequence database required

Protein ID by Sequence Query
Protein ID by Sequence Query

Mass values combined with amino acid sequence or composition data

Figure 1. Principle of matching peptide sequence tags to a pre-provided sequence. The upper chain of amino acids represents the peptide sequence as measured by MS/MS (from Table 1 in this example), and the lower chain represents amino acids in the sequence database that the tag is compared to. Note that the partial sequence divides the peptide into three regions. The added mass \( m_1 \) of the residues in region 1, together with the N-terminus, is a match criterion as is the added mass in region three, \( m_2 \). In region 2, the sequence is known. Furthermore, it can be required that the peptide obey the cleavage condition of the proteolytic enzyme, marked by K/R for trypsin. The left pointing arrow indicates that both search directions may have to be considered.


Picking out a good tag

1489.430 tag(650.213, GWSV, 1079.335)

From: John Cottrell, Matrix Science
Performing a Sequence Query

Sequence Query Result
Sequence Query Result

Sequence Query with Different Possibilities

1890.2 tag(1004.1, Seq, 1548.5)

Seq =LSADTG or LSAMG or LSAFG
=LSA[DT|M|F]G
Error Tolerant Sequence Tag

A sequence tag can match to a peptide despite there being an unsuspected modification or point mutation by allowing the mass values to ‘float’. For example, take the peptide GVQVETISP GDGR, MH+ = 1314.7 and the (b ion) sequence tag:
1314.7 tag(614.3,TISP,911.5)

If there was an unsuspected modification on the N-terminal side of the tag, which increased the mass by 100, this would affect both the fragment ion mass values in tandem. The tag interpreted from the spectrum would become:
1414.7 tag(714.3,TISP,1011.5)

On the other hand, if the unsuspected modification was on the C-terminal side of the tag, or if the fragment ions were y series ions, the fragment ion mass values would be unchanged, and the interpreted tag would be:
1414.7 tag(614.3,TISP,911.5)

Error Tolerant Sequence Query
Sequence Tag

Rapid search times

Error tolerant
• Match peptide with unknown modification or SNP

Requires interpretation of spectrum
• Usually manual, hence not high throughput

Tag has to be called correctly
• Although ambiguity is OK
2060.78 tag(977.4,[Q|K][Q|K][Q|K]EE,1619.7)

From: John Cottrell, Matrix Science

Sequence Tag Programs on the Web

Sequence Tag / Sequence Homology
MultiTag

GutenTag

MS-Blast
Shevchenko, A., et al., Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time of flight mass spectrometry and BLAST homology searching, Analytical Chemistry 73 1917-1926 (2001)

FASTS, FASTF

OpenSea

CIDentify

Sequence queries can be extremely powerful. These references are a good starting point if you are interested in learning more about the potential of combining mass and sequence information.

From: John Cottrell, Matrix Science
De Novo Sequencing
### Dipeptide MH+ ion Table
(by Karl Clauser)

<table>
<thead>
<tr>
<th>Gln</th>
<th>Asn</th>
<th>Ser</th>
<th>Pro</th>
<th>Val</th>
<th>Thr</th>
<th>Cys</th>
<th>Leu</th>
<th>Ile</th>
<th>Met</th>
<th>His</th>
<th>Arg</th>
<th>Asp</th>
<th>Glu</th>
<th>Gly</th>
</tr>
</thead>
<tbody>
<tr>
<td>110</td>
<td>114</td>
<td>106</td>
<td>103</td>
<td>113</td>
<td>114</td>
<td>111</td>
<td>112</td>
<td>117</td>
<td>119</td>
<td>119</td>
<td>118</td>
<td>119</td>
<td>120</td>
<td>118</td>
</tr>
</tbody>
</table>

Gly + Gly = 114.043 u and Asn = 114.043 u
Ala + Gly = 128.059 u and Glu = 128.059 u and Lys = 128.095 u
Gly + Val = 156.090 u and Arg = 156.101 u
Ala + Asp = Glu + Gly = 186.064 u and Trp = 186.079 u
Ser + Val = 196.100 u and Trp = 186.079 u

---

**References:**
- David W. Speicher, The Wistar Institute
**De Novo Sequencing**

(M+2H)²⁺ is most commonly observed when using tryptic peptide

Doubly charged spectra give easiest spectra to interpret manually

David W. Speicher, The Wistar Institute

---

**De Novo Sequencing - Application**

Analyze high quality spectra left after database matching

Obtain sequence for proteins not in database

Obtain experience with MS/MS data for manual validation of database matching

David W. Speicher, The Wistar Institute
De Novo Sequencing - Factors

1. Neutral losses
   \( \text{H}_2\text{O} (-18) - y \) and \( b \)-ions from S,T,D,E
   \( \text{NH}_3(-17) - y \) and \( b \)-ions from R,K,(N,Q)
   \( \text{C}=\text{O} (-28) - a \)-ions from small \( b \)-ions

2. Low mass cut-off on ion traps = 28% of precursor ion

3. Bond on N-terminal size of P, very labile
   Large \( y \)-ion ending in P
   Internal fragments starting at \( P^+ \) several a.a.
   Minor in ion trap, substantial on triple quads and QToF's

4. Immonium ions
   low mass ions indicative of presence of a cer
   BUT lack of observed immonium ion does not
   that a.a. - not useful for ion traps due to low

5. Isobaric amino acids & amino acid combinations
   - L, I
   - Q, K < 0.03 Da

David W. Speicher, The Wistar Institute

Residue Massed of Amino Acids

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Monoisotopic mass (^a)</th>
<th>Average mass</th>
<th>Immonium ion mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>Gly or G</td>
<td>57.02</td>
<td>57.05</td>
<td>-</td>
</tr>
<tr>
<td>Alanine</td>
<td>Ala or A</td>
<td>71.04</td>
<td>71.08</td>
<td>-</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser or S</td>
<td>87.03</td>
<td>87.08</td>
<td>-</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro or P</td>
<td>97.05</td>
<td>97.12</td>
<td>72</td>
</tr>
<tr>
<td>Valine</td>
<td>Val or V</td>
<td>99.07</td>
<td>99.13</td>
<td>-</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr or T</td>
<td>101.05</td>
<td>101.11</td>
<td>-</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys or C</td>
<td>103.01</td>
<td>103.14</td>
<td>-</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu or L</td>
<td>113.08</td>
<td>113.16</td>
<td>101</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile or I</td>
<td>113.08</td>
<td>113.16</td>
<td>86</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asp or N</td>
<td>134.04</td>
<td>114.10</td>
<td>-</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Glu or Q</td>
<td>128.06</td>
<td>128.17</td>
<td>128.17</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys or K</td>
<td>128.09</td>
<td>128.17</td>
<td>128.17</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Gla or E</td>
<td>129.04</td>
<td>129.12</td>
<td>129.12</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met or M</td>
<td>137.13</td>
<td>137.12</td>
<td>104</td>
</tr>
<tr>
<td>Histidine</td>
<td>His or H</td>
<td>137.06</td>
<td>137.14</td>
<td>110</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe or F</td>
<td>147.07</td>
<td>147.18</td>
<td>147.18</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg or R</td>
<td>156.10</td>
<td>156.19</td>
<td>156.19</td>
</tr>
<tr>
<td>Carbamidomethyl cysteine</td>
<td>Ccarc</td>
<td>160.03</td>
<td>160.3</td>
<td>160.3</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr or Y</td>
<td>163.05</td>
<td>163.18</td>
<td>136</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Try or W</td>
<td>186.08</td>
<td>186.21</td>
<td>159</td>
</tr>
</tbody>
</table>

\[(M+2H^+)2^+= (\sum \text{Residue Mass} + 18 +2)/2\]
Common Immonium Ions

- b-ion series → $b_1$ m/z=?, $b_{14}$ m/z=?
- y-ion series → $y_1$ m/z=?, $y_{14}$ m/z=?

<table>
<thead>
<tr>
<th>#</th>
<th>b</th>
<th>b''</th>
<th>b'</th>
<th>b**</th>
<th>b0</th>
<th>b''''</th>
<th>Seq</th>
<th>y</th>
<th>y''</th>
<th>y0</th>
<th>y''''</th>
<th>y''+</th>
<th>y0+</th>
<th>y''''+</th>
<th>m/z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>129.10</td>
<td>65.05</td>
<td>112.08</td>
<td>56.54</td>
<td></td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>228.17</td>
<td>114.59</td>
<td>211.14</td>
<td>106.08</td>
<td>V</td>
<td>1511.84</td>
<td>756.43</td>
<td>1494.82</td>
<td>747.91</td>
<td>1493.83</td>
<td>747.42</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>325.22</td>
<td>163.12</td>
<td>308.20</td>
<td>154.60</td>
<td>P</td>
<td>1412.77</td>
<td>706.89</td>
<td>1395.75</td>
<td>698.38</td>
<td>1394.76</td>
<td>697.89</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>453.28</td>
<td>227.14</td>
<td>436.26</td>
<td>218.63</td>
<td>Q</td>
<td>1315.72</td>
<td>658.36</td>
<td>1298.70</td>
<td>649.85</td>
<td>1297.71</td>
<td>649.36</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>553.35</td>
<td>276.68</td>
<td>535.32</td>
<td>268.17</td>
<td>V</td>
<td>1187.66</td>
<td>594.34</td>
<td>1170.64</td>
<td>585.82</td>
<td>1169.65</td>
<td>585.33</td>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>639.38</td>
<td>320.19</td>
<td>622.36</td>
<td>311.68</td>
<td>S</td>
<td>1088.59</td>
<td>544.86</td>
<td>1071.57</td>
<td>536.29</td>
<td>1070.58</td>
<td>535.80</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>740.43</td>
<td>370.72</td>
<td>723.40</td>
<td>362.21</td>
<td>T</td>
<td>1001.56</td>
<td>501.28</td>
<td>984.54</td>
<td>492.77</td>
<td>983.55</td>
<td>492.25</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>837.48</td>
<td>419.25</td>
<td>820.46</td>
<td>410.73</td>
<td>P</td>
<td>900.51</td>
<td>450.76</td>
<td>883.49</td>
<td>442.25</td>
<td>882.50</td>
<td>441.76</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>918.56</td>
<td>469.77</td>
<td>901.59</td>
<td>461.26</td>
<td>T</td>
<td>803.46</td>
<td>402.23</td>
<td>786.44</td>
<td>393.72</td>
<td>785.45</td>
<td>393.23</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1051.61</td>
<td>526.31</td>
<td>1034.59</td>
<td>517.80</td>
<td>L</td>
<td>702.41</td>
<td>351.71</td>
<td>685.39</td>
<td>341.20</td>
<td>684.40</td>
<td>341.71</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1150.68</td>
<td>585.85</td>
<td>1133.66</td>
<td>576.33</td>
<td>V</td>
<td>589.33</td>
<td>295.37</td>
<td>572.30</td>
<td>286.66</td>
<td>571.32</td>
<td>286.16</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1279.73</td>
<td>648.47</td>
<td>1262.70</td>
<td>631.85</td>
<td>E</td>
<td>490.26</td>
<td>245.63</td>
<td>473.24</td>
<td>237.12</td>
<td>472.25</td>
<td>236.63</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1377.79</td>
<td>698.90</td>
<td>1360.77</td>
<td>681.39</td>
<td>Y</td>
<td>361.22</td>
<td>181.11</td>
<td>344.19</td>
<td>172.60</td>
<td>343.21</td>
<td>172.11</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1465.83</td>
<td>753.42</td>
<td>1448.80</td>
<td>724.90</td>
<td>S</td>
<td>262.35</td>
<td>131.59</td>
<td>245.12</td>
<td>123.07</td>
<td>244.14</td>
<td>122.57</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>R</td>
<td>175.12</td>
<td>88.06</td>
<td>158.09</td>
<td>79.55</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Residue Massed of Amino Acids

(N-term)KVPQVSTPTLVEVSR (15 Residues, MW=1638.93)