Primer on Proteomics

Prepared by Choon-Peng Chng, Bioinformatics Institute

March 2003
About Proteomics

- This is the study of the proteome, which is simply the set of proteins created from a genome.
- We study multi-protein complexes.
- Interested in identifying components of proteins (AA sequence) via experiments.
Proteomic Analysis
Flowchart

Protein mixture → separation → Proteins

Peptide mixture → separation → Peptides

Proteins → digestion → Identification

Peptides → digestion → MS analysis → MS data → Database search algorithms
Mass Spectrometry (MS)

Figure from [1].
Types of MS

Two primary types:

- MALDI-TOF: matrix-assisted laser desorption ionization (source) - time of flight (analyzer).
  - Primarily used to measure peptide masses.
- ESI-Tandem: electrospray ionization (source) - multistage mass analyzers such as ion-trap.
  - Obtain sequence data.

Please refer to the Appendix for a primer on Mass Spectrometry techniques.
Data analysis

- The MS data is used, together with specialized software, to identify peptides and proteins.

- Two main approaches:
  - Peptide mass finger-printing with MALDI data.
  - Protein sequence identification with ESI data.

- Method 2 is more useful as we can have a situation whereby different peptides in database gives same m/z value as our peptide produced by cleavage.

- ESI produces multiply-charge peptide ions.

- Tandem MS can induce peptide fragmentation from which we can derive sequence information such patterns in MS-MS spectra.
Peptide fragmentation nomenclature

Peptide ions created from the source collides with neutral gas atoms in the tandem mass spectrometer and the absorbed kinetic energy induces fragmentation.

The most significant cleavages are along the peptide backbone.

The y-ion is a fragment where the positive charge is retained on the C-terminus of the original peptide ion.

We wish to obtain doubly-charged ions such that both y- and b- ions are produced. The other cleavage sites (a, z, c, x) require higher energies to produce.

Source: Chapt 8 Fig 2 [1]
Peptide fragmentation example
Residue masses

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>One-letter code</th>
<th>Average residue mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>G</td>
<td>57.05</td>
</tr>
<tr>
<td>Alanine</td>
<td>A</td>
<td>71.08</td>
</tr>
<tr>
<td>Serine</td>
<td>S</td>
<td>87.08</td>
</tr>
<tr>
<td>Valine</td>
<td>V</td>
<td>99.13</td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>103.14</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tryptophan</td>
<td>W</td>
<td>186.21</td>
</tr>
</tbody>
</table>
Annotated MS-MS spectrum

Source: Chapter 8 fig. 4 of [1]
de novo sequence interpretation

After getting the MS-MS spectrum, we can sit down with a calculator and a table of AA residue masses and try to interpret the b- and y- ions.

Once we get the sequence, do a BLAST search against a protein sequence DB to identify the protein.

This can be done in 15mins to hours, depending on one’s experience level.

Clearly this is too slow for large-scale protein identification, as is the case with proteomics.
The manual interpretation was bypassed by using algorithms to correlate spectra with peptide sequences in databases, without interpreting each spectrum individually. These are known as data-reduction software, such as Sequest. If we assume that the algorithm is accurate enough, then the only limiting factors are the spectra quality and the database completeness/accuracy. Furthermore, databases are growing rapidly due to the many genome projects.
However, if the sequence of the analyzed protein does not exist in the database, then a correct match cannot be made.

Also, RNA alternative splicing as well as posttranslational modifications will give us problems.

Hence we still need de novo sequencing!!
TRASH - automated data reduction

- Thorough High Resolution Analysis of Spectra by Horn.
- Proposed by David Horn, Roman A. Zubarev and Fred W. McLafferty of the Department of Chemistry, Baker Lab, Cornell University in 1999.
- Dr Newman Sze of the Proteomics group in GIS has collaborations with the Baker Lab and had been using TRASH in Cornell and would like to make THRASH run faster. Hence our involvement to see if we can speed up THRASH.
TRASH features

- Provides integrated data analysis of multiply charged ions of a complete ESI mass spectrum.
- Identifies the isotopic clusters, resolves those that overlap and subject these to automated charge determination algorithm and least-squares isotopic abundance distribution-fitting algorithm.
- The idea is to remove human intervention for the de novo method.
Algorithm description

Figure 2. Flowchart summary of the automated program.
Main stages

- **Input parameters:**
  - Desired m/z range in the spectrum to be analyzed
  - Signal-to-noise (S/N) threshold for initiation of the isotopic abundance-fitting procedure (algorithm initiation threshold)
  - Range of expected charge states in the spectrum.
  - Optional: minimum reliability value (RL) for a successful least-squares fit.
- **Charge determination.**
  - Find best fit of calculated and experimental abundances to identify the monoisotopic peak.
  - Repeat process until all possible isotopic clusters in the m/z range are characterized.
Mass determination

- MS instruments measure the mass/charge ratio. Hence we need to determine the correct number of charges an ion has.

- Example: If $m/z = 556.1$ and charge state=2, then $m=556.1 \times 2=1112.2$. However, to get the $m$(neutral), we need to subtract off 2.

- Formula: $m/z = (M+nH+)/n$

- In practice, we do not know the charge states.

- We can use the fact that Carbon atoms have 2 main isotopes, C-12 and C-13, which exists naturally in organic compounds in some ratio.

- A carbon atom in a peptide which is present as C-13 will have a mass+1 larger than the all C-12 version (also known as the...
Monoisotopic peak

- This is usually the first peak in the multi-charge envelope. All other peaks that contains C-13 will be lower and shifted to it's right-side.

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

Average mass is calculated based on the average atomic weights, measured from the centroid of the isotopic envelope.
Effect of resolution

Mass resolution: ratio of the mass of the peak over the peak width, taken as the full width at half maximum intensity (fwhm).
The larger the better, since the peak would be narrower!

Molecular weight determined corresponds to the average peptide mass using low resolution and the monoisotopic mass using higher resolution.
**Charge-determination**

- When the resolution and S/N is high, the reciprocal of the spacing between adjacent isotopic peaks gives z.

- However, for high charge states, we’re unable to achieve the required accuracy/resolution.

- Patterson/Fourier transform methods from pattern recognition is used for such cases, which uses the entire isotopic envelope. They are expected to give maxima that corresponds to repetitive spacing between isotopic peaks.

- Also applicable for overlapping clusters of different z as maxima are independently determined. Note that each isotopic cluster has its own charge state z.

- If the RL value is < 90%, a backup method employing least-squares (see later; more time consuming) is used to improve the accuracy.
Monoisotopic mass determination

- The peak created by the all-monoisotopic species (all 12C, 1H, 14N, 16O, 32S) is best for the determination of the peptide mass.

- However, for large molecules this species has negligible abundance, say only < 4% for bovine ubiquitin (8.6 kDa). This is because it’s more probable to have multiple isotopes here.

- We can estimate the monoisotopic mass by the average mass, which is found via fitting the measured spectra with a model isotopic distributions.

- The charge-state is obtained from the previous step.

- Possible problem: b- or y- ions?
Averagine

This is a model amino acid with elemental components occurring at frequencies deduced from the PIR database.

Formerly people assumed a uniform distribution, such that sulfur-containing AA like Cysteine and Methionine appear 5% of the time (1/20). However, it was found that the percentages were only 1.9% and 2.3%.

The deduced model AA (averagine) is:

- $C_{4.9384} H_{7.7583} N_{1.3577} O_{1.4773} S_{0.0417}$
- with molecular mass of 111.1254 Da.

To construct a poly-averagine molecule (peptide) of known mass, say horse myoglobin with mass 16,951 Da, we get $(16951/111)=152.5$ units of averagine needed. We then multiply by # of atoms of each element to get $C_{753} H_{1206} N_{207} O_{225} S_{6}$. 
Use of Least-squares

- Using the above formula, a theoretical isotopic profile is created.

- This is then overlapped with the experimental peaks using a least-squares procedure in order to determine the most abundant (tallest) peak.

- Note: For low mass isotopic clusters, the most abundant peak is also the monoisotopic peak.

Figure 6. From the Figure 5 spectrum, least-squares fit assignment of the most abundant peak using (circles) only the five isotopic peaks S/N > 3, leading to a 1 Da error, and using (squares) the peaks of abundance >20% of that of the most abundant peak, leading to the correct assignment.
Least-squares charge-determination

- The least-squares fitting is also used as a backup method to improve on the charge determination should the Patterson/FT method fails.

- We step through the range of possible charge-states and choose \( z \) with largest reliability (RL) value.

- Once we have identified the correct charge and \( m/z \), we can get the mass of the fragment. Say for the first cluster, it's \( 991.5 \times 6 = 5949 - 3 \) (?)

   kill 2 birds with 1 stone!!
References


Appendix: Mass Spectrometry Techniques
How the MALDI source works

- Sample to be analyzed is mixed with a chemical matrix, typically containing a small organic molecule that absorbs light at a specific wavelength.

- This admixture is then spotted onto a small plate/slide and allowed to evaporate.

- Evaporation of the solvent allows the formation of a crystal lattice, which contains the peptide.

- This target is then placed into the source, equipped with a laser.

- The matrix chemical absorbs the laser light and transfers energy to the peptides, which are ejected out into the gas phase.
Ions produced by MALDI

- Both positive and negative ions are produced.
- For proteins and peptides, we’re interested in the positive ions, formed by accepting a proton. Most ions are singly charged.
- Example: A peptide of mass 1032 becomes an ion of m/z 1033 and represented by [M+H]+.
TOF Mass Analyzers

- The Time-Of-Flight instrument measures the time between the ion entering the analyzer to the moment it hits the detector.
- The speed of the ions is directly proportional to their m/z values.
- Earliest TOF analyzers operated in this "linear" mode.

Figure: Schematic representation of a MALDI-TOF MS (Page 58 of Ref[1])
TOF Mass Analyzers (cont’d)

- However, resolution is poor, which is the ability to distinguish between ions of slightly different m/z values.
- This is due to variation in speeds of ions with same m/z.
- Modern MALDI-TOF comes with a reflectron, which acts to focus ions of identical m/z such that they reach the detector at the same time.
Peaks due to isotopic ions (12C vs 13C, say) are resolved.

Another approach is the use of pulsed-laser ionization with delayed extraction.

The goal is to let the ions get a "fair start" such that ions of same m/z arrive at the detector together.

We average over spectra obtained from many laser pulses (10-100).

Same improvement of spectral resolution obtained as for reflectron.
Pros and Cons of MALDI

- Pros
  - Very easy to operate.
  - Compatible with robotic sample preparation devices for high-throughput proteomics.
  - Accurate enough for peptide mass measurements.
  - Very sensitive, using femtomole down to attomole (10^-18 mole) sources.
Pros and Cons of MALDI

- **Cons**
  - Cannot provide sequence information, like from peptide ion fragmentation.
  - Greatly affected by quality of the sample, which can be contaminated.
ESI source

- Peptides or proteins to be analyzed are in aqueous form.
- Makes use of the property that functional groups in peptides acquire different charge at different pH.
- We usually work in acidic pH where protonation of amines confer net positive charge to peptides, which helps in fragmentation.
- ESI unique in that it produces multiply-charged ions. Many peptides bare multiple proton-accepting sites.
- This also helps to keep ions within mass (actually m/z) range of the ESI analyzers which have a more limited range than TOF.
- Hence, ESI mass spectrum contains a “multi-charge envelope” that contains all the charge states of the protein.
How the ESI source works

Figure: Schematic representation of an ESI source (Page 68 of Ref[1])
Deconvolution technique

- ESI produces multiply charged peptide ions, so we don’t get 1 peak in the spectrum for the peptide but a “multicharge envelope” of signals.

- Charge deconvolution algorithms can reduce the envelope (which looks Gaussian) to a signal peak.

- Combines peaks of the same mass but different charge, taking advantage of the fact that charges have integer values.
Deconvolution technique
Tandem Mass Analyzers

- Three common types:
  - Triple quadrupole ("triple quad")
  - Ion-trap
  - Quadruple TOF

- They all perform the same task: select a single m/z species.

- The ion is then subjected to collision-induced dissociation (CID), which breaks it into fragment ions as well as neutral fragments.

- Tandem (or MS-MS) refers to the chain of MS analyses that the ions are subjected to, either through 3 quadrupoles analyzers or cycles in the same ion-trap.

- Please refer to "Introduction to Proteomics" for details.
Fourier-Transform MS

- The Fourier transform ion cyclotron resonance mass spectrometer (known as FT-ICR or more commonly, FTMS) is analogous to an ion-trap.
- The mass analyzer employs a powerful magnetic field and FT algorithms to detect all ions in the trap simultaneously.
- Able to achieve spectacular resolution.
- Potentially very powerful to analytical proteomics but very expensive and temperamental.