Mass spectrometry in proteomics

Haixu Tang
School of Informatics and Computing
Indiana University, Bloomington

Modified from: www.bioalgorithms.info

Outline

- Proteomics & Mass spectrometry
- Application of MS/MS in proteomics
  - Protein sequencing and identification by mass spectrometry
    - Protein Identification via Database Search (SPC & spectral alignment)
    - De Novo Peptide Sequencing (Spectrum graph)
  - Hybrid
    - Identifying Post Translationally Modified (PTM) Peptides
  - (Quantitative proteomics)
    - identifying proteins that are differentially abundant

Entering the era of human omics

- Genomic and other omic approach will become a part of routine healthcare practice
- Genomics: revealing static genetic information in somatic cells; may alter in tumor cells
- Omics approaches: high-throughput monitoring of cellular systems at the genome-scale
  - Transcriptomics: dynamic alteration of the forms and abundances of proteins
  - Glycomics: dynamic alteration of glycansynthesis
  - Epigenomics: dynamic alteration of chromatin (on which the genome is physically located) status
  - Metabolomics: dynamic alteration of metabolic reactions
  - Immunomics: dynamic alteration of immune response

Mike Snyder (Stanford): integrated Personal Omics Profiling (iPOP)

- An analysis that combines genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles from a single individual over a 14 month period;
  - Every half a month to two months
- Uncover extensive, dynamic changes in diverse molecular components and biological pathways across healthy and diseased conditions;
- Revealed various medical risks, including type II diabetes
- Can be used to interpret healthy and disease states by connecting genomic information with additional dynamic omics activity;

The Dynamic Nature of the Proteome

- The proteome of the cell is changing
  - Various extra-cellular, and other signals activate pathways of proteins.
  - A key mechanism of protein activation is post-translational modification (PTM)
  - These pathways may lead to other genes being switched on or off
  - Mass spectrometry is key to probing the proteome and detecting PTMs

Dynamic proteomic profiling

- Identify proteins in a complex sample (e.g., human blood samples)
  - Shotgun approach (bottom-up); Top-down approach
- Quantify these proteins in samples from various conditions (e.g., disease vs. healthy, dynamic sampling from multiple time points of the same individual)
- Identify post-translation modifications (PTMs) of proteins (phosphorylations, acetylations, glycosylations, etc)
  - Different PTMs are dynamic with different time scales
An analytical technique for the determination of the elemental composition of a sample or molecule

Ion source: ESI (electrospray ionization), MALDI (matrix-assisted laser desorption/ionization)

Mass analyzer: separate the ions according to their mass-to-charge ratio, e.g., TOF (time-of-flight)

Mass Spectrometry (MS)

Shotgun proteomics

Protein identification
Inferring which proteins are present in the sample
A probabilistic point of view: the likelihood of each protein (in the database) being in the sample

Peptide-spectrum matchings (PSMs)

Industrial standard: report identified peptides by controlled false discovery rate (FDR) – the forward-decoy strategy
- Search both the target and a decoy database (e.g. the reverse protein database)
- Use the peptide-spectrum matches (PSMs) in decoy database to estimate the false PSMs in the target database
- FDR = # decy PSMs / # target PSMs
- Can be used for any search engine or scoring model

Protease, e.g. trypsin, break protein into peptides.
A Tandem Mass Spectrometer further breaks the peptides down into fragment ions and measures the mass of each piece.
Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones.
Mass Spectrometer measure mass/charge ratio of an ion.
Breaking Proteins into Peptides

\[ \text{protein} \rightarrow \text{peptides} \]

Tandem Mass Spectrometry

- Tandem Mass Spectrometry (MS/MS): mainly generates partial N- and C-terminal peptides
- Chemical noise often complicates the spectrum.
- Represented in 2-D: mass/charge axis vs. intensity axis

Protein Identification with MS/MS

MS/MS Peptide Identification:

N- and C-terminal Peptides
N- and C-terminal Peptides

Peptide Shotgun Sequencing

Reconstruct peptide from the set of masses of fragment ions (mass-spectrum)

Issue 1: Spectrum Consists of Different Ion Types

Because peptides can be broken in several places.

Issue 2: Noise and Missing Peaks

The peaks in the mass spectrum:

- Prefix and Suffix Fragments.
- Fragments with neutral losses (-H₂O, -NH₃)
- Noise and missing peaks.

De Novo vs. Database Search

Database Search

De Novo

Database of known peptides

AVGELTK
Peptide Identification Problem (Database Search)

**Goal:** Find a peptide from the database with maximal match between an experimental and theoretical spectrum.

**Input:**
- S: experimental spectrum
- database of peptides
- Δ: set of possible ion types
- m: parent mass

**Output:**
- A peptide of mass m from the database whose theoretical spectrum matches the experimental S spectrum the best

---

Peptide Identification by Database Search

- Compare experimental spectrum with theoretical spectra of database peptides to find the best fit
- The match between two spectra is the number of masses (peaks) they share (Shared Peak Count or SPC)
- In practice mass-spectrometrists use the weighted SPC that reflects intensities of the peaks
- Match between experimental and theoretical spectra is defined similarly
- To find the peptide with theoretic spectrum that is most similar to the real spectrum

---

Peptide Sequencing Problem (De Novo)

**Goal:** Find a peptide with maximal match between an experimental and theoretical spectrum.

**Input:**
- S: experimental spectrum
- Δ: set of possible ion types
- m: parent mass

**Output:**
- A peptide with mass m, whose theoretical spectrum matches the experimental S spectrum the best

---

De novo Peptide Sequencing

**Building Spectrum Graph**

- How to create vertices (from masses)
- How to create edges (from mass differences)
- How to score paths
- How to find best path

---

Intensity

Mass/Charge (M/Z)

a is an ion type shift in b
MS/MS Spectrum (Ion Types Unknown & With Noise)

Some Mass Differences between Peaks Correspond to Amino Acids

Knowing Ion Types

- Some masses correspond to fragment ions, others are just random noise
- Knowing ion types $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$ lets us distinguish fragment ions from noise
- We can learn ion types $\delta_i$ and their probabilities $q_i$ by analyzing a large test sample of annotated spectra.

Vertices of Spectrum Graph

- Masses of potential N-terminal peptides
- Vertices are generated by reverse shifts corresponding to ion types $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$
- Every N-terminal peptide can generate up to $k$ ions $m \cdot \delta_1, m \cdot \delta_2, \ldots, m \cdot \delta_k$
- Every mass $s$ in an MS/MS spectrum generates $k$ vertices $V(s) = \{s + \delta_1, s + \delta_2, \ldots, s + \delta_k\}$ corresponding to potential N-terminal peptides
- Vertices of the spectrum graph: \{initial vertex\}$\cup V(s_1) \cup V(s_2) \cup \ldots \cup V(s_m) \cup \text{terminal vertex}\$

Reverse Shifts

- Two peaks $b \cdot H_2O$ and $b$ are given by the Mass Spectrum
- With a $+H_2O$ shift, if two peaks coincide that is a possible vertex.
**Edges of Spectrum Graph**

- Two vertices with mass difference corresponding to an amino acid \( A \):
  - Connect with an edge labeled by \( A \)
- Gap edges for di- and tri-peptides

**Paths**

- Path in the labeled graph spell out amino acid sequences
- There are many paths, how to find the correct one?
- We need **scoring** to evaluate paths

**Path Score**

- \( p(P, S) = \text{probability that peptide } P \text{ produces spectrum } S = \{s_1, s_2, \ldots, s_q\} \)
- \( p(P, s) = \text{the probability that peptide } P \text{ generates a peak } s \)
- **Scoring** = computing probabilities
- \( p(P, S) = \prod_{s \in S} p(P, s) \)

**Finding Optimal Paths in the Spectrum Graph**

- For a given MS/MS spectrum \( S \), find a peptide \( P' \) maximizing \( p(P, S) \) over all possible peptides \( P \):
  \[ p(P', S) = \max_P p(P, S) \]
- Peptides = paths in the spectrum graph
- \( P' = \text{the optimal path in the spectrum graph} \)

**De Novo vs. Database Search: A Paradox**

- The database of all peptides is huge \( \approx O(2^n) \).
- The database of all known peptides is much smaller \( \approx O(10^n) \).
- However, *de novo* algorithms can be much faster, even though their search space is much larger!
- A database search scans all peptides in the **database of all known peptides** search space to find best one.
- *De novo* eliminates the need to scan database of all peptides by modeling the problem as a graph search.
- **But De novo** sequencing is still not very accurate!

**Sequencing of Modified Peptides**

De novo peptide sequencing is invaluable for identification of unknown proteins:

However, *de novo* algorithms are designed for working with high quality spectra with good fragmentation and without modifications.

Another approach is to compare a spectrum against a set of known spectra in a database.
Protein identification problem: a probabilistic formulation

\[ P(P_i^1 = 1 | S_j) = \prod_j P(S_j | q_j = 1) P(q_j = 1 | P_i = 1) P(P_i = 1) \]

Challenges:
1. Is peptide detectability predictable?
2. How to assess PSMs?
3. How to model degenerate peptides?

Basic Bayesian model

\[ \text{MAP solution: } \min_{x_1, x_2, ..., x_m} \sum_l P(S_l | x_1, x_2, ..., x_m) \]

Huge protein configuration space (2^n putative solutions)

Extending the probabilistic formulation to protein inference problem

- Protein configuration graph
  - Include not only identified peptides, but also non-identified peptides
  - Edge weights represent the prior probabilities of observing a peptide in a shotgun proteomics platform (i.e., peptide detectabilities)

Probabilistic approach to protein inference problem

Extreme case: multiple proteins sharing the same (single) identified peptide

Probabilistic formulation: Bayesian models

Combinatorial formulation: minimum missed peptides

Incorporation of peptide detectability: the probability that the peptide will be observed in a standard sample analyzed by a standard proteomics routine
From protein identification to protein quantification

\[ A_{ijz} = r_{ijz} \cdot q_i + \epsilon_{ijz} \]

Peak area of peptide \( j \) charged \( c \) from protein \( i \)

Quantity (abundance) of protein \( i \)

Peptide response rate

Background noise

Assumption: 1) considering only unique (non-redundant) peptides; 2) errors model, Log-normal/Poisson: \( \sigma^2_{ijz} \propto r_{ijz}^2 \); Gaussian: \( \sigma^2_{ijz} \propto r_{ijz} \)

Post-Translational Modifications

Proteins are involved in cellular signaling and metabolic regulation.

They are subject to a large number of biological modifications.

Almost all protein sequences are post-translationally modified and 200 types of modifications of amino acid residues are known.

Examples of Post-Translational Modification

Phosphorylation

Glycosylation

Post-translational modifications increase the number of "letters" in amino acid alphabet and lead to a combinatorial explosion in both database search and de novo approaches.

Identification of Peptides with Mutations:

Challenge

Very similar peptides may have very different spectra (so SPC won’t work!)

Goal: Define a notion of spectral similarity that correlates well with the sequence similarity.

If peptides are a few mutations/modifications apart, the spectral similarity between their spectra should be high.

Similar Peptides with Different Spectra

<table>
<thead>
<tr>
<th>No mutations</th>
<th>1 mutation</th>
<th>2 mutations</th>
</tr>
</thead>
</table>

Problem: SPC diminishes very quickly as the number of mutations increases. (Only a small portion of correlations between the spectra is captured by SPC.)

Search for Modified Peptides:

Virtual Database Approach

Yates et al., 1995: an exhaustive search in a virtual database of all modified peptides.

Exhaustive search leads to a large combinatorial problem, even for a small set of modifications types.

Problem (Yates et al., 1995). Extend the virtual database approach to a large set of modifications.
Spectral Convolution

\[ S_2 \otimes S_1 = \{ s_2 - s_1 : s_1, s_2 \in S_1 \times S_2 \} \]

Number of pairs \( s_1, s_2 \in S_2 \) with \( s_2 - s_1 = x \):

\[ (S_2 \otimes S_1)(x) \]

The shared peaks count (SPC peak):

\[ (S_2 \otimes S_1)(0) \]

Spectral Comparison: Difficult Case

\[ S = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\} \]

Which of the spectra

\[ S' = \{10, 20, 30, 40, 50, 55, 65, 75, 85, 95\} \]

or

\[ S'' = \{10, 15, 30, 35, 50, 55, 70, 75, 90, 95\} \]

fits the spectrum \( S \) the best?

SPC: both \( S' \) and \( S'' \) have 5 peaks in common with \( S \).

Spectral Convolution: reveals the peaks at \( 0 \) and \( 5 \).

Limitations of the Spectrum Convolutions

Spectral convolution does not reveal that spectra \( S \) and \( S' \) are similar, while spectra \( S \) and \( S'' \) are not.

Clumps of shared peaks: the matching positions in \( S' \) come in clumps while the matching positions in \( S'' \) don’t.

This important property was not captured by spectral convolution.

Shifts

\[ A = \{ a_1 < \ldots < a_n \} : \] an ordered set of natural numbers.

A shift \((i, \Delta)\) is characterized by two parameters, the position \( i \) and the length \( \Delta \).

The shift \((i, \Delta)\) transforms

\[ \{ a_1, \ldots, a_n \} \]

into

\[ \{ a_{i+\Delta}, \ldots, a_{i+n-1} \} \]
Shifts: An Example

The shift \((i, \Delta)\) transforms \(\{a_1, \ldots, a_n\}\) into \(\{a_1, \ldots, a_i\Delta, a_{i+\Delta}, \ldots, a_n\Delta\}\)

e.g.

\[
\begin{align*}
10 & 20 & 30 & 40 & 50 & 60 & 70 & 80 & 90 & \\
10 & 20 & 30 & 35 & 45 & 55 & 65 & 75 & 85 & \\
10 & 20 & 30 & 35 & 45 & 55 & 62 & 72 & 82 & \\
\end{align*}
\]

Spectral Alignment Problem

- Find a series of \(k\) shifts that make the sets \(A = \{a_1, \ldots, a_n\}\) and \(B = \{b_1, \ldots, b_n\}\) as similar as possible.

- \(k\)-similarity between sets

- \(D(k)\) - the maximum number of elements in common between sets after \(k\) shifts.

Representing Spectra in 0-1 Alphabet

- Convert spectrum to a 0-1 string with 1s corresponding to the positions of the peaks.

Comparing Spectra=Comparing 0-1 Strings

- A modification with positive offset corresponds to inserting a block of 0s
- A modification with negative offset corresponds to deleting a block of 0s
- Comparison of theoretical and experimental spectra (represented as 0-1 strings) corresponds to a (somewhat unusual) edit distance/alignment problem where elementary edit operations are insertions/deletions of blocks of 0s
- Use sequence alignment algorithms!

Spectral Alignment vs. Sequence Alignment

- Manhattan-like graph with different alphabet and scoring.
- Movement can be diagonal (matching masses) or horizontal/vertical (insertions/deletions corresponding to PTMs).
- At most \(k\) horizontal/vertical moves.

Use of \(k\)-Similarity

SPC reveals only \(D(0) = 3\) matching peaks.
Spectral Alignment reveals more hidden similarities between spectra: \(D(1) = 5\) and \(D(2) = 8\) and detects corresponding mutations.
Protein Identification

- We can detect peptides from mass spectra by database search or de novo approaches
- Homologous proteins

References

- Mass spectrometry-based proteomics, Nature 422:198, 2003
- Applying mass spectrometry-based proteomics to genetics, genomics and network biology, Nature Reviews Genetics 10, 617-627, 2009