CSE182-L12

Mass Spectrometry
Peptide identification
Mass-Charge ratio

• The X-axis is not mass, but \( \frac{M+Z}{Z} \)
  – \( Z=1 \) implies that peak is at \( M+1 \)
  – \( Z=2 \) implies that peak is at \( \frac{M+2}{2} \)
    • \( M=1000, Z=2, \) peak position is at 501

• Quiz: Suppose you see a peak at 501. Is the mass 500, or is it 1000?
Isotopic peaks

- Ex: Consider peptide SAM
- Mass = 308.12802
- You should see:
  - Instead, you see
    - 308.13
    - 310.13
All atoms have isotopes

- Isotopes of atoms
  - O\textsubscript{16,18}, C-12,13, S\textsubscript{32,34}…
  - Each isotope has a frequency of occurrence
- If a molecule (peptide) has a single copy of C-13, that will shift its peak by 1 Da
- With multiple copies of a peptide, we have a distribution of intensities over a range of masses (Isotopic profile).
- How can you compute the isotopic profile of a peak?
Isotopes

- C-12 is the most common. Suppose C-13 occurs with probability 1%
- EX: SAM
  - Composition: C11 H22 N3 O5 S1
- What is the probability that you will see a single C-13?

Note that C,S,O,N all have isotopes. Can you compute the isotopic distribution?

\[
\binom{11}{1} \cdot 0.01 \cdot (0.99)^{10}
\]
Isotope Calculation

- Denote:
  - $N_c$: number of carbon atoms in the peptide
  - $P_c$: probability of occurrence of C-13 (~1%)
- Then

\[
\Pr[\text{Peak at } M] = \binom{N_c}{0} p_c^0 (1 - p_c)^{N_c}
\]

\[
\Pr[\text{Peak at } M + 1] = \binom{N_c}{1} p_c^1 (1 - p_c)^{N_c - 1}
\]

\[
N_c = 50
\]

\[
N_c = 200
\]
Isotope Calculation Example

- Suppose we consider Nitrogen, and Carbon
- $N_N$: number of Nitrogen atoms
- $P_N$: probability of occurrence of N-15
- $P(\text{peak at } M)$
- $P(\text{peak at } M+1)$?
- $P(\text{peak at } M+2)$?

Pr[Peak at $M$] = \[\binom{N_C}{0}p_c^0(1-p_c)^{N_C}\binom{N_N}{0}p_N^0(1-p_N)^{N_N}\]

Pr[Peak at $M+1$] = \[\binom{N_C}{1}p_c^1(1-p_c)^{N_C-1}\binom{N_N}{0}p_N^0(1-p_N)^{N_N} + \binom{N_C}{0}p_c^0(1-p_c)^{N_C}\binom{N_N}{1}p_N^1(1-p_N)^{N_N-1}\]

How do we generalize? How can we handle Oxygen ($O-16,18$)?
General isotope computation

• Definition:
  – Let $p_{i,a}$ be the abundance of the isotope with mass $i$ Da above the least mass
  – Ex: $P_{0,C}$: abundance of C-12, $P_{2,O}$: O-18 etc.
  – Let $N_a$ denote the number of atoms of amino-acid $a$ in the sample.

• Goal: compute the heights of the isotopic peaks. Specifically, compute $P_i = \text{Prob}\{M+i\}$, for $i=0,1,2\ldots$
We define the characteristic polynomial of a peptide as follows:

\[ \phi(x) = P_0 + P_1 x + P_2 x^2 + P_3 x^3 + \ldots \]

\[ \phi(x) \] is a concise representation of the isotope profile.
Consider a single carbon atom. What is its characteristic polynomial

\[ \phi(x) = P_0 + P_1 x + P_2 x^2 + P_3 x^3 + \ldots \]
\[ = p_{0,c} + p_{1,c} x \]
• Suppose carbon was the only atom with an isotope C-13. In a peptide, if we have $N_c$ carbon atoms, what is the isotope profile?

$$\phi(x) = P_0 + P_1 x + P_2 x^2 + P_3 x^3 + \ldots$$

$$= \binom{N_c}{0} p_{0,c}^N (1 - p_{0,c})^0 + \binom{N_c}{1} p_{0,c} (1 - p_{0,c})^1 x$$

$$= (p_{0,c} + p_{1,c} x)^{N_c}$$
Consider a molecule with one carbon atom, and one oxygen atom. What is the isotope profile?

\[ \phi(x) = P_0 + P_1x + P_2x^2 + P_3x^3 + \ldots \]

\[ = (p_{0,c} + p_{1,c}x)(p_{0,O} + p_{2,O}x^2) \]
General isotope computation

- Definition:
  - Let $p_{i,a}$ be the abundance of the isotope with mass $i$ Da above the least mass
  - Ex: $P_{0,C}$: abundance of C-12, $P_{2,O}$: O-18 etc.
- Characteristic polynomial

- $\text{Prob}\{M+i\}$: coefficient of $x^i$ in $\phi(x)$ (a binomial convolution)

$$\phi(x) = \prod_a \left( p_{0,a} + p_{1,a}x + p_{2,a}x^2 + \cdots \right)^{N_a}$$
Isotopic Profile Application

- In DxMS, hydrogen atoms are exchanged with deuterium
- The rate of exchange indicates how buried the peptide is (in folded state)
- Consider the observed characteristic polynomial of the isotope profile $\phi_{t_1}, \phi_{t_2}$, at various time points. Then

- The estimates of $p_{1,H}$ can be obtained by a deconvolution
- Such estimates at various time points should give the rate of incorporation of Deuterium, and therefore, the accessibility.

$$\phi_{t_2}(x) = \phi_{t_1}(x)(p_{0,H} + p_{1,H})^{N_H}$$

Not in Syllabus

CSE182
Quiz

- How can you determine the charge on a peptide?
  - Difference between the first and second isotope peak is $1/Z$

- Proposal:
  - Given a mass, predict a composition, and the isotopic profile
  - Do a ‘goodness of fit’ test to isolate the peaks corresponding to the isotope
  - Compute the difference
Ion mass computations

- Amino-acids are linked into peptide chains, by forming peptide bonds
  - Residue mass
    - \( \text{Res.Mass}(aa) = \text{Mol.Mass}(aa) - 18 \)
    - (loss of water)
Peptide chains

- \( \text{MolMass(SGFAL)} = \text{resM}(S) + \ldots \text{res}(L) + 18 \)
M/Z values for b/y-ions

- Singly charged b-ion = ResMass(prefix) + 1
- Singly charged y-ion = ResMass(suffix)+18+1
- What if the ions have higher units of charge?
De novo interpretation

• Given a spectrum (a collection of b-y ions), compute the peptide that generated the spectrum.
• A database of peptides is not given!
• Useful?
  – Many genomes have not been sequenced
  – Tagging/filtering
  – PTMs
De Novo Interpretation: Example

Ion Offsets

\[ b = P + 1 \]
\[ y = S + 19 = M - P + 19 \]
Computing possible prefixes

- We know the parent mass $M=401$.
- Consider a mass value 88
- Assume that it is a b-ion, or a y-ion
- If b-ion, it corresponds to a prefix of the peptide with residue mass $88-1 = 87$.
- If y-ion, $y = M - P + 19$.
  - Therefore the prefix has mass
    - $P = M - y + 19 = 401 - 88 + 19 = 332$
- Compute all possible Prefix Residue Masses (PRM) for all ions.
Putative Prefix Masses

- Only a subset of the prefix masses are correct.
- The correct mass values form a ladder of amino-acid residues

<table>
<thead>
<tr>
<th></th>
<th>Prefix Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>M=401</td>
<td>b</td>
</tr>
<tr>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>145</td>
<td>144</td>
</tr>
<tr>
<td>147</td>
<td>146</td>
</tr>
<tr>
<td>276</td>
<td>275</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S</th>
<th>G</th>
<th>E</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>87</td>
<td>144</td>
<td>273</td>
</tr>
</tbody>
</table>
Spectral Graph

- Each prefix residue mass (PRM) corresponds to a node.
- Two nodes are connected by an edge if the mass difference is a residue mass.
- A path in the graph is a de novo interpretation of the spectrum
• Each peak, when assigned to a prefix/suffix ion type generates a unique prefix residue mass.
• Spectral graph:
  – Each node $u$ defines a putative prefix residue $M(u)$.
  – $(u,v)$ in $E$ if $M(v) - M(u)$ is the residue mass of an a.a. (tag) or 0.
  – Paths in the spectral graph correspond to a interpretation
Re-defining *de novo* interpretation

- Find a subset of nodes in spectral graph s.t.
  - 0, M are included
  - Each peak contributes at most one node (interpretation)(*).
  - Each adjacent pair (when sorted by mass) is connected by an edge *(valid residue mass)*
  - An appropriate objective function (ex: the number of peaks interpreted) is maximized

![Diagram of a spectral graph showing nodes 87 and 144 connected by an edge, with mass labels S, G, E, K, and M.]
Two problems

- **Too many nodes.**
  - Only a small fraction are correspond to b/y ions (leading to true PRMs) (learning problem)
- **Multiple Interpretations**
  - Even if the b/y ions were correctly predicted, each peak generates multiple possibilities, only one of which is correct. We need to find a path that uses each peak only once (algorithmic problem).
  - In general, the forbidden pairs problem is NP-hard
Too many nodes

• We will use other properties to decide if a peak is a b-y peak or not.
• For now, assume that $\delta(u)$ is a score function for a peak $u$ being a b-y ion.
Multiple Interpretation

• Each peak generates multiple possibilities, only one of which is correct. We need to find a path that uses each peak only once (algorithmic problem).
• In general, the forbidden pairs problem is NP-hard
• However, The b,y ions have a special non-interleaving property
• Consider pairs \((b_1,y_1), (b_2,y_2)\)
  – If \(b_1 < b_2\), then \(y_1 > y_2\)
- If we consider only b,y ions, ‘forbidden’ node pairs are non-intersecting,
- The de novo problem can be solved efficiently using a dynamic programming technique.
The forbidden pairs method

- Sort the PRMs according to increasing mass values.
- For each node \( u \), \( f(u) \) represents the forbidden pair.
- Let \( m(u) \) denote the mass value of the PRM.
- Let \( \delta(u) \) denote the score of \( u \).
- Objective: Find a path of maximum score with no forbidden pairs.
D.P. for forbidden pairs

- Consider all pairs u,v
  - \( m[u] \leq M/2, \ m[v] > M/2 \)

- Define \( S(u,v) \) as the best score of a forbidden pair path from
  - 0->u, and v->M

- Is it sufficient to compute \( S(u,v) \) for all u,v?
D.P. for forbidden pairs

• Note that the best interpretation is given by

\[
\max_{((u,v)\in E)} S(u,v)
\]
D.P. for forbidden pairs

- Note that we have one of two cases.
  1. Either \( u > f(v) \) (and \( f(u) < v \))
  2. Or, \( u < f(v) \) (and \( f(u) > v \))
- Case 1.
  - Extend \( u \), do not touch \( f(v) \)

\[
S(u, v) = \max_{u':(u',u) \in E} \left( S(u', v) + \delta(u) \right)
\]
The complete algorithm

for all u /*increasing mass values from 0 to M/2 */
for all v /*decreasing mass values from M to M/2 */
    if (u < f[v])
        \[ S[u, v] = \max_{(v, w) \in E \atop w \neq f(u)} \left( S[u, w] + \delta(v) \right) \]
    else if (u > f[v])
        \[ S[u, v] = \max_{(w, u) \in E \atop w \neq f(v)} \left( S[w, v] + \delta(u) \right) \]
If (u,v) \in E
    /* maxI is the score of the best interpretation */
    maxI = \max \{ maxI, S[u,v] \}
EXTRA SLIDES
De Novo: Second issue

• Given only b,y ions, a forbidden pairs path will solve the problem.
• However, recall that there are MANY other ion types.
  – Typical length of peptide: 15
  – Typical # peaks? 50-150?
  – #b/y ions?
  – Most ions are “Other”
    • a ions, neutral losses, isotopic peaks….
De novo: Weighting nodes in Spectrum Graph

- Factors determining if the ion is b or y
  - Intensity (A large fraction of the most intense peaks are b or y)
  - Support ions
  - Isotopic peaks
De novo: Weighting nodes

- A probabilistic network to model support ions (Pepnovo)

\[
\text{Score}(m, S) = \log \frac{P_{\text{CID}}(I|m, S)}{P_{\text{RAND}}(I|m, S)}
\]

**Figure 1.** Probabilistic network for the CID fragmentation model of doubly charged tryptic peptides measured in an ion trap mass spectrometer. Three different types of relations are modeled in this network: (1) correlations between fragment ions (regular arrows); (2) dependencies due to the relative position of the cleavage site in the peptide (dashed arrows); (3) influence of flanking amino acids to the cleavage site (bold arrows).
De Novo Interpretation Summary

- The main challenge is to separate b/y ions from everything else (weighting nodes), and separating the prefix ions from the suffix ions (Forbidden Pairs).
- As always, the abstract idea must be supplemented with many details.
  - Noise peaks, incomplete fragmentation
  - In reality, a PRM is first scored on its likelihood of being correct, and the forbidden pair method is applied subsequently.
- In spite of these algorithms, de novo identification remains an error-prone process. When the peptide is in the database, db search is the method of choice.
The dynamic nature of the cell

- The proteome of the cell is changing
- Various extra-cellular, and other signals activate pathways of proteins.
- A key mechanism of protein activation is PT modification
- These pathways may lead to other genes being switched on or off
- Mass Spectrometry is key to probing the proteome
Biol. Data analysis: Review
Other static analysis is possible
A Static picture of the cell is insufficient

- Each Cell is continuously active,
  - Genes are being transcribed into RNA
  - RNA is translated into proteins
  - Proteins are PT modified and transported
  - Proteins perform various cellular functions
- Can we probe the Cell dynamically?
  - Which transcripts are active?
  - Which proteins are active?
  - Which proteins interact?
Counting transcripts

- cDNA from the cell hybridizes to complementary DNA fixed on a ‘chip’.
- The intensity of the signal is a ‘count’ of the number of copies of the transcript.
Quantitation: transcript versus Protein Expression

Our Goal is to construct a matrix as shown for proteins, and RNA, and use it to identify differentially expressed transcripts/proteins
Gene Expression

- Measuring expression at transcript level is done by micro-arrays and other tools.
- Expression at the protein level is being done using mass spectrometry.
- Two problems arise:
  - Data: How to populate the matrices on the previous slide? (‘easy’ for mRNA, difficult for proteins)
  - Analysis: Is a change in expression significant? (Identical for both mRNA, and proteins).
- We will consider the data problem here. The analysis problem will be considered when we discuss micro-arrays.
MS based Quantitation

- The intensity of the peak depends upon
  - **Abundance**, ionization potential, substrate etc.
- We are interested in abundance.
- Two peptides with the same abundance can have very different intensities.
- **Assumption:** *relative* abundance can be measured by comparing the ratio of a peptide in 2 samples.
Quantitation issues

• The two samples might be from a complex mixture. How do we identify identical peptides in two samples?
• In micro-array this is possible because the cDNA is spotted in a precise location? Can we have a ‘location’ for proteins/peptides
LC-MS based separation

- As the peptides elute (separated by physiochemical properties), spectra is acquired.
LC-MS Maps

- A peptide/feature can be labeled with the triple (M,T,I):
  - monoisotopic M/Z, centroid retention time, and intensity
- An LC-MS map is a collection of features
Time scaling: Approach 1 (geometric matching)

- Match features based on M/Z, and (loose) time matching. Objective $\Sigma f(t_1-t_2)^2$
- Let $t_2' = a t_2 + b$. Select $a, b$ so as to minimize $\Sigma f(t_1-t_2')^2$
Geometric matching

- Make a graph. Peptide a in LCMS1 is linked to all peptides with identical m/z.
- Each edge has score proportional to $t_1/t_2$
- Compute a maximum weight matching.
- The ratio of times of the matched pairs gives $a$.
- Rescale and compute the scaling factor $T_{M/Z}$.
Approach 2: Scan alignment

- Each time scan is a vector of intensities.
- Two scans in different runs can be scored for similarity (using a dot product)

\[
S_{1i} = 10 \ 5 \ 0 \ 0 \ 7 \ 0 \ 0 \ 2 \ 9
\]

\[
S_{2j} = 9 \ 4 \ 2 \ 3 \ 7 \ 0 \ 6 \ 8 \ 3
\]

\[
M(S_{1i}, S_{2j}) = \sum_k S_{1i}(k) S_{2j}(k)
\]
Scan Alignment

- Compute an alignment of the two runs
- Let $W(i,j)$ be the best scoring alignment of the first $i$ scans in run 1, and first $j$ scans in run 2

$$W(i,j) = \max \left\{ W(i-1,j-1) + M[S_{1i}, S_{2j}] \\
W(i-1,j) + \ldots \\
W(i,j-1) + \ldots \right\}$$

- Advantage: does not rely on feature detection.
- Disadvantage: Might not handle affine shifts in time scaling, but is better for local shifts
Chemistry based methods for comparing peptides
ICAT

- The reactive group attaches to Cysteine
- Only Cys-peptides will get tagged
- The biotin at the other end is used to pull down peptides that contain this tag.
- The X is either Hydrogen, or Deuterium (Heavy)
  - Difference = 8Da
ICAT

- ICAT reagent is attached to particular amino-acids (Cys)
- Affinity purification leads to simplification of complex mixture
Differential analysis using ICAT
ICAT issues

- The tag is heavy, and decreases the dynamic range of the measurements.
- The tag might break off
- Only Cysteine containing peptides are retrieved Non-specific binding to strepdavidin
Serum ICAT data

MA13_02011_02_ALL01Z3I9A* Overview (exhibits 'stack-ups')
• Instead of pairs, we see entire clusters at 0, +8, +16, +22.

• ICAT based strategies must clarify ambiguous pairing.
ICAT problems

- Tag is bulky, and can break off.
- Cys is low abundance
- MS$_2$ analysis to identify the peptide is harder.
**SILAC**

- A novel stable isotope labeling strategy
- Mammalian cell-lines do not ‘manufacture’ all amino-acids. Where do they come from?
- Labeled amino-acids are added to amino-acid deficient culture, and are incorporated into all proteins as they are synthesized
- No chemical labeling or affinity purification is performed.
- Leucine was used (10% abundance vs 2% for Cys)
SILAC vs ICAT

- Leucine is higher abundance than Cys
- No affinity tagging done
- Fragmentation patterns for the two peptides are identical
  - Identification is easier

Ong et al. MCP, 2002
Incorporation of Leu-d3 at various time points

- Doubling time of the cells is 24 hrs.
- Peptide = VAPEEHPVLLTEAPLNPK
- What is the charge on the peptide?
Quantitation on controlled mixtures
Identification

- MS/MS of differentially labeled peptides
Peptide Matching

- Computational: Under identical Liquid Chromatography conditions, peptides will elute in the same order in two experiments.
  - These peptides can be paired computationally
- SILAC/ICAT allow us to compare relative peptide abundances in a single run using an isotope tag.