Mass Spectrometry
Peptide identification
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 $i=0,1,2...$
Characteristic polynomial

• 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
Characteristic polynomial computation

Suppose carbon was the only atom with an isotope C-13.

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

\[ = \binom{N_C}{0} p_{0,c}^N (1 - p_{0,c})^0 + \binom{N_C}{1} p_{0,c} (1 - p_{0,c})^1 x \]

\[ = (p_{0,c} + p_{1,c} x)^{N_C} \]
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

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

- $\text{Prob}\{M+i\}$: coefficient of $x^i$ in $\phi(x)$ (a binomial convolution)
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 $\phi_{t1}$, $\phi_{t2}$, at various time points. Then

$$\phi_{t2}(x) = \phi_{t1}(x)(p_{0,H} + p_{1,H})^{N_H}$$

• 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.

Not in Syllabus
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
  - (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

\[ \begin{align*}
0 & \quad 88 & 145 & 274 & 402 & \text{b-ions} \\
420 & 333 & 276 & 147 & 0 & \text{y-ions}
\end{align*} \]

**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.
Spectral Graph

- 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 an 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
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 \(y_1 > y_2\)
Non-Intersecting Forbidden pairs

- 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 \( f(v) \)

\[
S(u,v) = \max_{u':(u',u) \in E \atop u' \neq f(v)} \left[ S(u',v) + \delta(u') \right]
\]
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])
    \[ S[u,v] = \max_{(v,w) \in E} S[u,w] + \delta(w) \]
  else if (u > f[v])
    \[ S[u,v] = \max_{(w,u) \in E} S[w,v] + \delta(w) \]
If (u,v)\in E
  /*maxI is the score of the best interpretation*/
  maxI = \max \{ \maxI, S[u,v] \}
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
What happens to the spectrum upon modification?

- Consider the peptide MSTYER.
- Either S, T, or Y (one or more) can be phosphorylated.
- Upon phosphorylation, the b-, and y-ions shift in a characteristic fashion. Can you determine where the modification has occurred?

If T is phosphorylated, $b_3, b_4, b_5, b_6,$ and $y_4, y_5, y_6$ will shift.
Effect of PT modifications on identification

- The shifts do not affect de novo interpretation too much. Why?
- Database matching algorithms are affected, and must be changed.
- Given a candidate peptide, and a spectrum, can you identify the sites of modifications
Db matching in the presence of modifications

• Consider MSTYER
• The number of modifications can be obtained by the difference in parent mass.
• If 1 phosphorylation, we have 3 possibilities:
  - MS*TYER
  - MST*YER
  - MSTY*ER
• Which of these is the best match to the spectrum?
• If 2 phosphorylations occurred, we would have 6 possibilities. Can you compute more efficiently?
Scoring spectra in the presence of modification

- Can we predict the sites of the modification?
- A simple trick can let us predict the modification sites?
- Consider the peptide ASTYER. The peptide may have 0, 1, or 2 phosphorylation events. The difference of the parent mass will give us the number of phosphorylation events. Assume it is 1.
- Create a table with the number of b,y ions matched at each breakage point assuming 0, or 1 modifications
- Arrows determine the possible paths. Note that there are only 2 downward arrows. The max scoring path determines the phosphorylated residue
Modifications

• Modifications significantly increase the time of search.
• The algorithm speeds it up somewhat, but is still expensive