CSE182-L8

Mass Spectrometry
Project Notes

- Implement a few tools for proteomics

  - C1: 11/2/04
    - Answer MS questions to get started, select project partner, select a project.

  - C2: 11/15/04 (All but web-team)
    - Plan of development, I/O format, I/O code
    - Supply I/O format and code to the web-team

  - C3: 11/22/04 (Web-team)
    - Build a mockup using code from all teams.

  - C4: 11/29/04
    - Final project demonstrations, followed by class presentations
Isotope Calculation

- Denote:
  - \( N_c \): number of carbon atoms in the peptide
  - \( P_c \): probability of occurrence of C-13
- Then

\[
\text{Pr[Peak at } M\text{]} = \frac{N_c}{0} P_c^0 \left(1 - P_c\right)^{N_c}
\]

\[
\text{Pr[Peak at } M + 1\text{]} = \frac{N_c}{1} P_c^1 \left(1 - P_c\right)^{N_c - 1}
\]
Isotope Calculation

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

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

$$Pr[\text{Peak at } M + 1] = \binom{N_C}{1} p_c^1 (1 - p_c)^{N_N} \binom{N_N}{0} p_N^0 (1 - p_N)^{N_N} + \binom{N_C}{0} p_c^0 (1 - p_c)^{N_N} \binom{N_N}{1} p_N^1 (1 - p_N)^{N_N}$$
De novo interpretation of mass spectra

- Recall that we discussed how peptides could be identified by scanning a database.
- What if the database did not contain the peptide of interest?
**De Novo Interpretation: Example**

**Ion Offsets**

- $b = P + 1$
- $y = S + 19 = M - P + 19$

<table>
<thead>
<tr>
<th>S</th>
<th>G</th>
<th>E</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>88</td>
<td>145</td>
<td>274</td>
</tr>
</tbody>
</table>

- b-ions

| 420 | 333 | 276 | 147 | 0 |

- y-ions

**M/Z**

- $b_1$, $y_1$, $y_2$, $b_2$
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>Prefix Mass</th>
<th>b</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>M=401</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>87</td>
<td>332</td>
</tr>
<tr>
<td>145</td>
<td>144</td>
<td>275</td>
</tr>
<tr>
<td>147</td>
<td>146</td>
<td>273</td>
</tr>
<tr>
<td>276</td>
<td>275</td>
<td>144</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.
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

(*)In general, finding paths using forbidden pairs is 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

- There may be many paths that avoid forbidden pairs.
- We choose a path that maximizes an objective function,
  - EX: the number of peaks interpreted
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.
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 $0\rightarrow u$, $v\rightarrow M$
- 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

\[
\max_{((u,v) \not\in E)} S(u, v)
\]
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 \( f(v) \)

\[
S(u, v) = \max_{(u', (u, u') \notin E \text{ or } u' \neq f(v))} S(u, u') + 1
\]
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])
            \[ S[u,v] = \max_{w \neq f(v)} S[w,v] + 1 \]
        else if (u < f[v])
            \[ S[u,v] = \max_{w \neq f(u)} S[u,w] + 1 \]
        If (u,v) \in E
            /*maxI is the score of the best interpretation*/
            \[ \text{maxI} = \max \{ \text{maxI}, S[u,v] \} \]
De Novo Interpretation Summary

- The main challenge is to separate the prefix ions from the suffix ions.
- The forbidden paths approach is a method to do just that.
- As always, the abstract idea must be supplemented with many details.
  - Multiple ion types, not just b, or y
  - 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.
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 ASTYER.
- 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 ASTYER
- The number of modifications can be obtained by the difference in parent mass.
- If 1 phosphorylation, we have 3 possibilities:
  - AS*TYER
  - AST*YER
  - ASTY*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
MS application: Protein-protein interaction

- Proteins combine to form functional complexes.
- An antibody is a special kind of protein that can recognize a specific protein.
- Use an antibody to recognize a protein in a complex. Isolate & Purify the complex that binds to the antibody.
- Identify all the proteins in the complex via mass spectrometry.
MS application: Protein Structure

- Use chemical cross-linkers to link spatially proximal residues.
- Denature and digest the protein. Identify the cross-linked peptides. This provides extra structural constraints which help predict structure.
Summary

- Mass Spectrometry data can be used to identify modified, and unmodified peptides.
- Additional applications in determining protein-protein interactions, protein structure etc.