Developing Algorithms for the Determination of Relative Peptide Abundances from LC/MS Data

Presented by: Arunima Ray, SUNY Geneseo

Jake Marcus, Anne Eaton, Melanie Kanter, Arunima Ray
Mentors: Shawn Cokus, Matteo Pellegrini, Parag Mallick, Roland Luethy
Proteomics and SFCAP

- Spielberg Family Center for Applied Proteomics (SFCAP)
  - Aims to develop and use proteomic tools to facilitate patient management decisions

**Proteomics**
- Large-scale study of proteins
- Proteins are the effectors of most biological functions
- The proteome is dynamic:
  - differs across cell types
  - changes over time based on external and internal factors
Individualized treatment

- Need to quantify the differences in proteomes
Data generation: LC/MS

- Start with a sample of proteins
- Protein sizes are varied:
  - Digest into fragments: **peptides**
- Separate by affinity to water
- Ionize
- Further separate by mass/charge
The data

List of Identifications

<table>
<thead>
<tr>
<th>Affinity to water</th>
<th>Mass/charge</th>
<th>Peptide</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>246</td>
<td>725.4</td>
<td>K.ACSQRPR.W</td>
<td>ADH</td>
</tr>
<tr>
<td>793</td>
<td>432.87</td>
<td>R.IGYADIK.W</td>
<td>EPO</td>
</tr>
<tr>
<td>1075</td>
<td>5367.91</td>
<td>K.LGANAILK.W</td>
<td>HB</td>
</tr>
</tbody>
</table>
Challenges

- Locate isotopes
- Identifications not centered
- Unknown spread along vertical axis
- Noise
Step 1: find associated feature

- Use affinity and charge information from identification
- Mass of the most abundant isotope
- Consider a region large enough to include the entire feature of interest
Step 2: consolidate information from isotopes

- Isotopes form parallel curves
- Select relevant mass/charge values, extract corresponding data
Step 2: consolidate information from isotopes
Step 3: limit along vertical axis (basic)

- Find highest peak
- Search along vertical axis until 4 out of 5 consecutive data points are below threshold
Step 3: limit along vertical axis (fingerprinting)

- Carbon in nature:
  - 98.93% in C-12 form
  - 1.07% in C-13 form
- For a peptide of n carbons:
  - Probability for all C-12 = 0.9893^n
  - Probability for 1 C-13, n-1 C-12 = n.(0.9893)^{n-1}(0.0107)
- Compute similarity score for intensities along mass/charge
- Graph for different affinity values
- Stop when scores below 10%
Step 4: fit a curve and quantify

- Need to model the distribution of intensities along vertical axis
- Gauss or gamma curve fit to data by nonlinear regression

\[ \text{Gamma, } R^2 = 0.98 \]
Evaluation

- 5 protein mix
  - (~250 peptides)
  - Different amounts in each sample

- 6 protein mix
  - (~150 peptides)
  - Same amount in every sample
Optimizing the algorithm
5 protein mix evaluations
6 protein mix evaluations
Outcomes

- Attempt with real data
  - The first (and only) run!

- Adaptable algorithm
- Modularity facilitates comparative analysis
- Addressed key challenges in quantification
- Investigated characteristics of features
Acknowledgments

- Institute for Pure and Applied Mathematics
- Spielberg Family Center for Applied Proteomics
- My teammates Jake, Anne and Melanie
- Our awesome mentors Shawn, Matteo, Roland and Parag.