SPECTRA-FIRST FEATURE ANALYSIS IN CLINICAL PROTEOMICS --- A CASE STUDY IN RENAL CANCER

GIW2016
Joint work with Limsoon Wong

Wilson Wen Bin Goh
The problem

• A lot of relevant signal is wasted when trying to map spectra to peptide/protein first (~70% of spectra is lost)
• Not all mapped spectra is interesting
• Why not first find differential spectra and map it to interesting peptides/proteins?
Renal cancer data (RC)

• 12 samples are run twice so that we have technical replicates over 6 normal and 6 cancer tissues

• $\sim$3200 observed proteins

• $\sim$30000 observed peptides (unique and non-unique)
A simple solution (1)

Condense retention time

Condense based on mz windows/bins
A simple solution (2)

Iterative deconvolution

Peptide matching based on mz coordinates
Most bin regions are dominated by only a few high-intensity spectra

Not likely that signal comes from multiple peptides in each bin
Rule-based feature-filtering

• **Rule 1**: If an MZ-bin has non-zero intensity values in more than half of the samples in class A and non-zero in more than half of the samples in class B, it is kept (for further filtering by Rule 2 below). Otherwise, it is discarded.

• **Rule 2**: For each class (A or B), the top 20% MZ-bins (ranked by summed intensity) supported by at least half of the samples in that class are kept.
Feature-selection based on MZ-Bin has good predictive power

<table>
<thead>
<tr>
<th>Group</th>
<th>MS1 filtered merged level 1 spectra</th>
<th>Single Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. significant features (0.05)</td>
<td>No. significant features (0.05)</td>
</tr>
<tr>
<td></td>
<td>CV Accuracy</td>
<td>CV p-value</td>
</tr>
<tr>
<td>1</td>
<td>82.00</td>
<td>0.83</td>
</tr>
<tr>
<td>2</td>
<td>130.00</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>179.00</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>234.00</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>220.00</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>158.00</td>
<td>0.75</td>
</tr>
<tr>
<td>7</td>
<td>96.00</td>
<td>0.92</td>
</tr>
<tr>
<td>8</td>
<td>166.00</td>
<td>0.83</td>
</tr>
<tr>
<td>9</td>
<td>230.00</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>98.00</td>
<td>0.92</td>
</tr>
<tr>
<td>mean</td>
<td>159.30</td>
<td>0.80</td>
</tr>
<tr>
<td>s.d.</td>
<td>167.89</td>
<td>0.80</td>
</tr>
<tr>
<td>COV</td>
<td>171.14</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Apparent loss of signal during peptide-to-protein transition

<table>
<thead>
<tr>
<th>Level</th>
<th>Sig_bins</th>
<th>pep_groups</th>
<th>proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>127</td>
<td>8249</td>
<td>2389</td>
</tr>
<tr>
<td>2</td>
<td>182</td>
<td>3018</td>
<td>1470</td>
</tr>
<tr>
<td>3</td>
<td>337</td>
<td>1044</td>
<td>688</td>
</tr>
</tbody>
</table>

B
Peptide quantitation is very varied

Most peptides are not reproducibly observed per protein!  If there is a lot of variability per peptide in the same protein, is it believable?
Using MZ-Bin to search for splice variants

• There must be at least 10 constituent unique peptides (to ensure reasonable coverage of the entire length of the protein);
• The peptides must be unambiguously mapped to the corresponding protein;
• At least 30% constituent peptides are over-expressed, i.e. $> 1.25$; and
• At least 30% constituent peptides are repressed, i.e. $< 0.8$. 
Using MZ-Bin to search for splice variants --- MAPT

Severe cancer
normals
Benign cancers?

MAPT, P10636

Severe cancer
normals
Benign cancers?

Exon 1: MAEPQFEFVMEHDHAGTYGLGDQMIDQEGYDQTDA^L
Exon 2: EPLAQFTPEDGSEEHPGSETDKSTPAE^L
Exon 3: DVTAPLYDEGAPKQAAAPHPTEIPEGTT^L
Exon 4: AEEAGIDTPSLEDEAAYHTQ^L
Exon 5: EPRSGLVQEGFLREPFGGLHLQMSMGPGAPLLEPGPREADTRPSGTGPEDEGGRHPPELLKHIQLGDLHIGEGPPPLKAGGGKRGSKKEVDEDDVDESSQIXPIDSAPAQDGPPPTAAREATSIQFPAEAGAIPVDFLSKVSIEIPASEPDPGFSQGRAKQGDAPL
Exon 6: ARMVSKSKDGTDSDKKAK^L
Exon 7: TSTRSSAKTLKNRCLSPKHPPTPSSDPPLQIOPSPAVCPEPSSPKYVSSVTSRGTSSGA
Exon 8: GDAGTICTIATPRGAPPGQKGAANATPRAPKTPPAKTPPSS^L
Exon 9: GEPKPSGDRGYYSGPSTGTGSRSTPSLTPFTRPPKAVVRTPPSXPSASKLRQ
Exon 10: VQIINKKLDLSVQSKGSKDNAKTVPGGS^L
Exon 11: VQIVYKPVDLKTSMKSGSILGNNHMK^L
Exon 12: GGGQVEVSXSKDLDFKDRVQSXDSKSIDNTHVPGGNNK^L
Exon 13: LETKHLTPXAKTDGAEMIVYKSPVSPIPGTDSRPHLSNVSSTGSDMVSDPQALTATLADYSASLAKQGL

^ Splice Junction
MAPT

- Associated with dementia
- Many known splice forms
- Reported as part of a severe RC predictive signature (down-regulated)
- But splice-variants associated with the down-regulation never reported before

- Down-regulated domains associated with severe phenotype maintain microtubule stability.
- This may have a role in impeding metastasis
- Consideration of specific peptides/domains may be more useful than considering the entire protein length
What have we learnt?

• Reverse proteomics analytical strategy is powerful
• A simple approach, MZ-Bin can lead to meaningful analytical outcome
• Beware of protein quantitation levels
• Peptide-based analysis can help us identify splice forms and get better insight on previously reported markers
References

• Goh & Wong. SPECTRA-FIRST FEATURE ANALYSIS IN CLINICAL PROTEOMICS --- A CASE STUDY IN RENAL CANCER. JBCB, In press
Acknowledgements

Professor Limsoon Wong
National University of Singapore