**Abstract**

Liquid chromatography coupled with two rounds of mass spectrometry (LC-MS/MS) applied in a technique known as 'shotgun proteomics', has proven effective as a means to capture a considerable portion of the protein complement of a biological sample. The technique often produces millions of mass spectra per experiment, however, approximately 50-75% of MS2 spectra remain unidentified, even as a good portion of these spectra are of high quality and likely peptide-derived. There are many possible reasons why these spectra may go unidentified including not having good enough database matches for spectra arising from biological phenotypes such as unknown post-translational modifications and simple nucleotide polymorphisms(SNPs). In addition, problems such as spectral clustering - a phenomenon where the isolation window for a peptide contains more than one distinct peak - is also known to negatively impact spectral library searching. Clustering these high-quality, unassigned (HQU) spectra together with their assigned counterparts, combined with a spectral purity analysis, may yield insights into the origins of these HQU spectra. This work combines clustering of mass spectra (using 3 distinct algorithms) with a spectral purity analysis. Moreover, the experimental design is leveraged by using the peptide intensities to identify unidentified spectra of possible biological relevance. The method is applied to an LC-MS2 dataset obtained from a circadian rhythm experiment in the plant species, *K. fedtschenkoi*. This plant is an important model species for the study of Circadian Acid Metabolism - a special adaptation of plants that inhabit areas with low water availability. Mining of this unstudied proteome resource may yield valuable insights into the metabolic changes that occur during the circadian rhythm of this plant.

**Introduction**

Clustering of mass spectra in proteomics refers to the grouping of mass spectra that have similar fragmentation patterns together into clusters. How these clusters are formed is an important aspect of the strategy to investigate unassigned spectra. These clustering algorithms used in the proteomics field are discussed here, all three having the basic following steps: Firstly, a similarity metric is used to compare all spectra to one another in a pairwise fashion. Secondly, from these similarities, these calculated pairwise distance relationships are used to cluster the spectra.

**Methods and Materials**

**LC and MS setup:**
All samples were analyzed on a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific) coupled with a Waters EASY-Spray LC liquid chromatography (LC) pump (Thermo Fisher Scientific). Peptides were separated on a 75 µm inner diameter, C18 reverse phase nanoscale capillary column (New Objective) with a 10 cm precolumn (packed in-house). For each sample, 2 µg of protein was loaded to buffer B (0:1% formic acid, 2% acetonitrile) and acetonitrile was increased to a final 5% (v:v) in buffer B in 10 min, followed by an increase in buffer B to 90% (v:v) in 15 min, another increase to 95% buffer B in 10 min, and finally increasing buffer B to 100% (v:v). The flow rate was kept at 200 nL/min. MS data was acquired with the Xcalibur software version 1.2.9, a method where the average scan rate was 1/2000 and the time between scans was 2.23 s. MS/MS spectra were acquired in 3 scans: one full scan, and two product scans (4 and 6 Da). Mass spectrometry was performed using a range of mass-to-charge ratios from 40 to 1,200 Da.

**Data analysis:**

- **MCluster** uses a normalized dot product. MClusterAutomatically uses a distance calculation relying on the rarity of experimental fragment peaks following the intuition that peaks shared by more than five spectra offer more evidence than peaks shared by a large number of spectra.
- **Spectra-cluster** uses a hypergeometric distribution to model the probability that the number of matched peaks occurred at random. The probability that the rank distribution of matched peaks occurred by chance is assessed using Kendall’s Tau correlation.

**Results**

Total number of scans: 2,232,718 produced with 1,845,503 passing noise filtering. 496,045 spectral assignments were made. Thus, 72% of spectra remain unassigned. Of the assigned spectra, 38,053 were unique and were used to infer 4,915 different proteins.

**Table 1. Summary of clustered scan produced by different algorithms**

<table>
<thead>
<tr>
<th>Spectral purity</th>
<th>MCluster</th>
<th>Spectra-cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lifetime</td>
<td>39%</td>
<td>44%</td>
</tr>
<tr>
<td>MS2 scan</td>
<td>39%</td>
<td>43%</td>
</tr>
<tr>
<td>MS1 scan</td>
<td>44%</td>
<td>41%</td>
</tr>
<tr>
<td>MS1 scan</td>
<td>39%</td>
<td>42%</td>
</tr>
<tr>
<td>MS2 scan</td>
<td>44%</td>
<td>41%</td>
</tr>
</tbody>
</table>

**Figure 1. Flow chart of the methodology used:**

1. Sample generation and LC-MS/MS analysis
2. Peptide spectral matching
3. Spectral clustering
4. Spectral purity calculation
5. Cluster similarity calculation
6. Cluster to Scan network construction

**Discussion**

Marastcy grouped the scans into the largest number of clusters followed by MCluster and then Spectra-cluster. All three algorithms were able to group nearly half of the scans (shown as clusters of size one). The bulk of the scans that could be grouped were formed part of clusters with less than 100 members. The cluster similarity analysis allowed that the clusters produced by Spectra-cluster were very similar to the clusters produced by the other algorithms. Marastcy and MCluster showed a large amount of correspondence in the composition of the clusters formed. Marastcy grouped the scans into the largest number of clusters followed by MCluster and then Spectra-cluster. Figure 4 shows a representative example of the structure for much of the data. It is centered around a cluster formed by Spectra-cluster. It has six spectra which are all grouped into different Marastcy and MClusters respectively. Only “Sample_031_Scan_30527” had a calculated purity score. Two plotted examples of the scans are presented.

**Conclusion**

The choice of clustering algorithm has a major impact on the grouping of scans for this particular dataset. Clusters that do share scans are very different in terms of size and the purity scores of the scans. No clear pattern between clustering behavior and spectral purity could be determined at this time since purity scores could only be determined for a third of the scans.