Pattern mining of mass spectrometry quality control data

Wout Bittremieux[1,2], Dirk Valkenborg[3], Aida Mrzic[1,2], Hanny Willems[3], Bart Goethals[1] and Kris Laukens[1,2]

Introduction

Mass spectrometry = identify proteins based on the mass distribution of their peptides ↔ results subject to a large variability
→ **Quality control metrics**

Previous research
- Univariate analysis [1]: Each metric individually ↔ does not take relationships between metrics into account
- Multivariate analysis [2]: Global analysis of all metrics simultaneously ↔ misses potentially interesting observations that are only expressed by a subset of metrics

As a result, **specialized pattern mining techniques** that take this duality into account can provide additional insights when analyzing quality control data.

Quality control data

Mass spectrometry data = **standardized quality control samples**
- Extensive frequency: periodically ran before, during, and after experimental samples
  → Detect problems as soon as possible
- Low complexity
  → Limited variability

Currently: quality control metrics for BSA samples run over the period of several years on a single Orbitrap Velos mass spectrometer.

Metrics = **identification-free metrics** (from QuaMeter [3])
- No (costly) peptide identification required
  → (Virtually) instantaneously available
- Does not depend on identification efficiency
  → Objective quality measures

Set of 44 quality control metrics involving different aspects of the mass spectrometer: chromatography, ionization, MS1, MS2, …

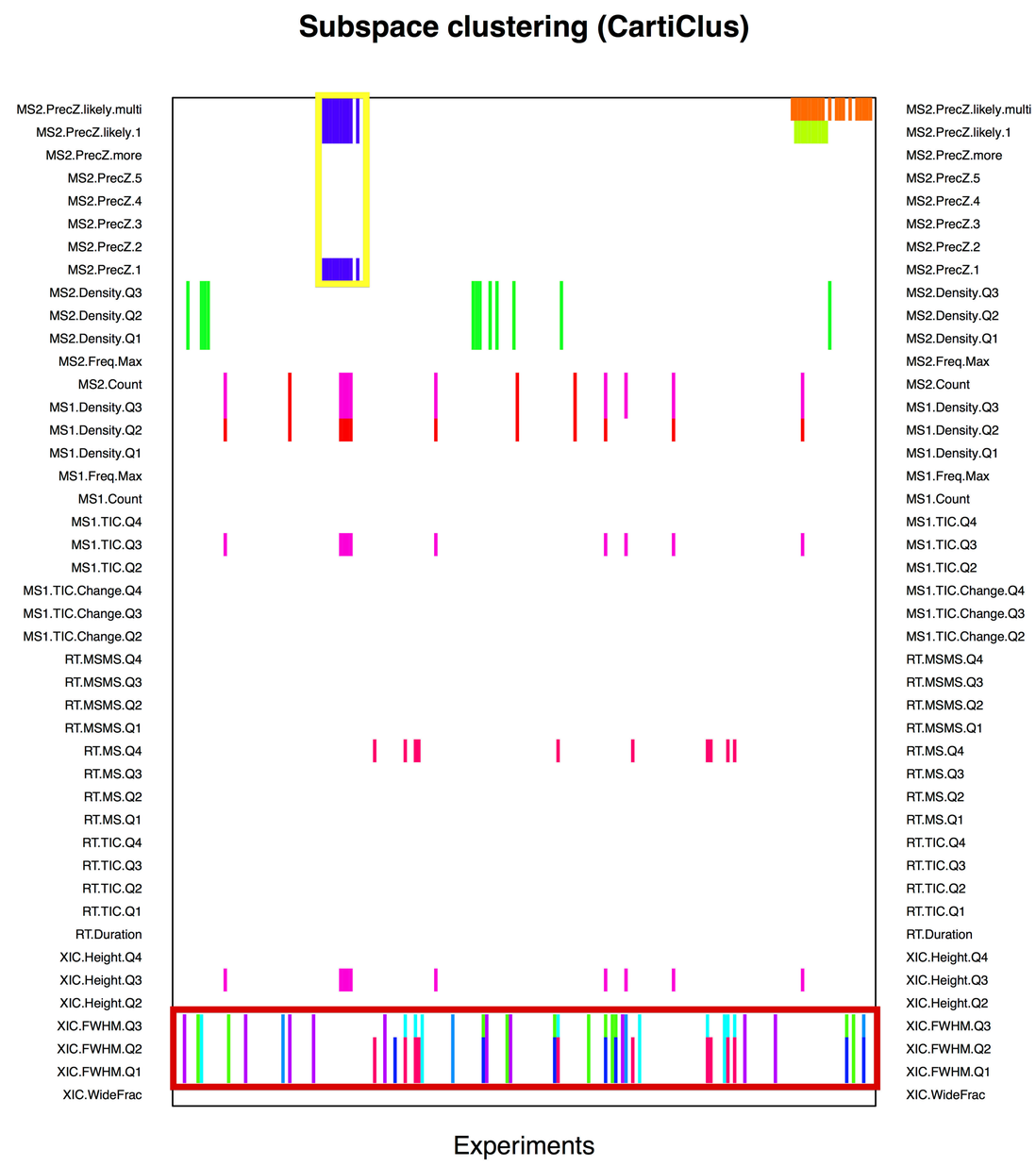

Figure 1: Subspace clustering can be used to detect patterns based on a subset of features. The subspaces found across multiple experiments are indicated by the same color.

Pattern mining

Prior analysis insufficient
- Univariate analysis: A single metric will be inadequate to detect problems expressed by the combination of multiple elements
- Multivariate analysis: Different sets of metrics are produced by a different generating mechanism (separate parts of a mass spectrometry experiment)

Solution = **subspace mining** algorithms
- Find a suitable subset of the original feature space by disregarding irrelevant dimensions
- Within each subspace: clustering, outlier detection, …

Subspace clustering [4] (Figure 1) = identify clusters of similar objects with respect to a subset of the attributes
- Highlighted in red: Several different subspace clusters concerning the full width at half maximum (FWHM), indicating different chromatography settings
- Highlighted in yellow: A potentially interesting cluster concerning the ionization of peptide fragments

Subspace outlier detection [5] (Figure 2) = detect highly deviating objects in any attribute combination
- Left: A collection of experiments that were detected as outliers, and where there was a known problem with the S-lens, which influenced the MS2 fragmentation
- Right: An example of potentially interesting outliers mainly exhibited by the chromatography-based metrics, which might indicate an unknown chromatography problem

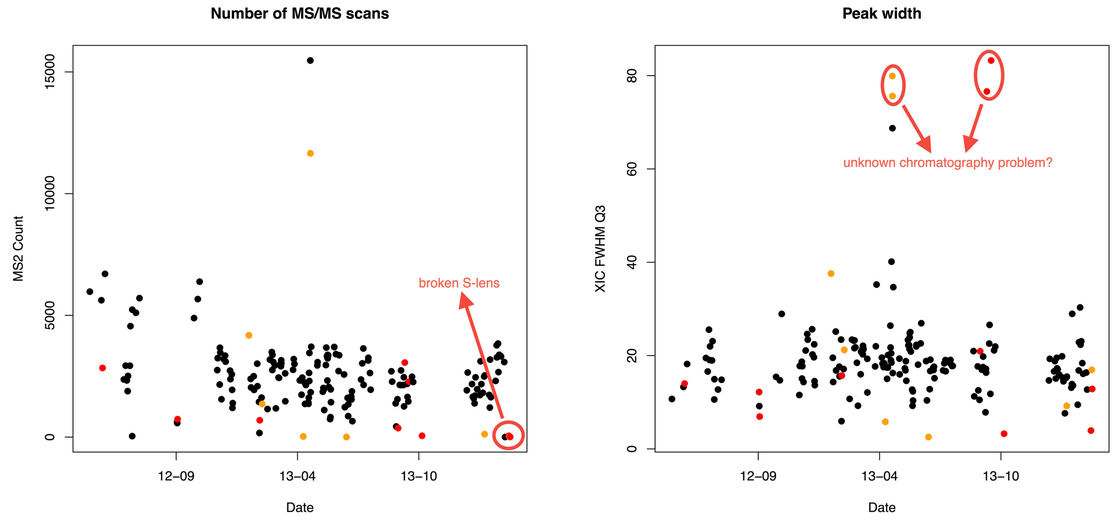

Figure 2: Subspace outlier mining can be used to detect outliers based on a subset of features. The top 10 ranked outliers are indicated in red, while the subsequent 10 ranked outliers are indicated in orange. Detected outliers are shown projected on a single dimension. Note that not all outliers are relevant in these dimensions.

Conclusion

Mass spectrometry experiments can be characterized by a broad set of quality control metrics. We have employed some specialized pattern mining techniques that take into account the specific properties of this high-dimensional data when looking for interesting patterns, such as subspace mining algorithms. Future work includes extending these initial approaches and relating the found patterns to experimental events.

References

1. Rudnick et al. (2010) *Molecular & Cellular Proteomics* **9**:225–241.
2. Wang et al. (2014) *Analytical Chemistry* **86**:2497–2509.
3. Ma et al. (2012) *Analytical Chemistry* **84**:5845–5850.
4. Aksehirli et al. (2013) *ICDM '13* 937–942.
5. Keller et al. (2012) *ICDE '12* 1037–1048.

wout.bittremieux@uantwerpen.be

Contact info

1 ADReM research group, Department of Mathematics and Computer Science, University of Antwerp, Belgium
2 Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp / Antwerp University Hospital, Belgium
3 Flemish Institute for Technological Research (VITO), Mol, Belgium

Universiteit Antwerpen