Mining mass spectrometry quality control data

Wout Bittremieux[1,2*], Pieter Kelchtermans[3,4,5], Dirk Valkenborg[4,5,6], Lennart Martens[3,4], Bart Goethals[1] and Kris Laukens[1,2]

[1]Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium; [2]Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp / Antwerp University Hospital, Antwerp, Belgium; [3]Department of Medical Protein Research, VIB, Ghent, Belgium, Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium; [4]Flemish Institute for Technological Research (VITO), Mol, Belgium; [5]CFP-CeProMa, University of Antwerp, Antwerp, Belgium; [6]I-BioStat, Hasselt University, Hasselt, Belgium
*wout.bittremieux@uantwerpen.be

Overview

Quality control is a very important aspect to assess the performance of a mass spectrometry experiment. Recently the **qcML** format has been proposed as a standard format for quality control information. In addition, the **jqcML** open-source Java API has been developed to work with qcML data. This standard data format and accessible API open up new possibilities to perform advanced **data mining** techniques, which can increase our understanding of complex mass spectrometry experiments.

Introduction

Because of the inherent complexity of mass spectrometry, the results of an experiment can be subject to a large **variability** (Figure 1). As a means of **quality control**, several qualitative metrics have been defined. However, these still suffer some **limiting factors**:

- **Compatibility**: Storing and communicating of quality control data is not standardized, limiting the dissemination along with experimental data;
- **Variability**: The data can be generated by software tools of different origins, with content and definitions varying for each tool.

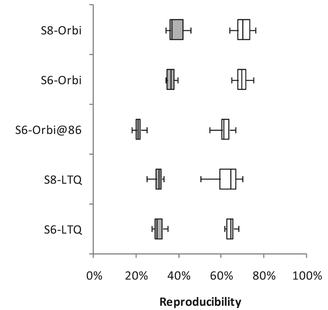

Figure 1: Reproducibility of identifications between different experiments on different instruments. Shaded boxes represent peptides, while white boxes represent proteins.[1]

qcML

In order to provide a pervasive and standardized means to report quality control information for mass spectrometry experiments, the **qcML standard**[2] has been developed. The qcML standard addresses these issues:

- **Compatibility**: XML-based file format (Figure 2; interchange format), and relational database (archival);
- **Variability**: Controlled vocabularies to unambiguously define terms.

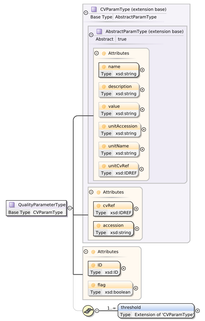

Figure 2: A quality metric as represented in the XML schema for the qcML standard version 0.0.8.

jqcML

jqcML[3] is an open-source **Java API** for working with qcML data:

- Complete **object model** to represent qcML data;
- The ability to work with data from **several sources** in a **uniform manner** (Figure 3).

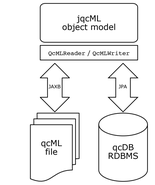

Figure 3: Simplified representation of the jqcML workflow.

Quality control data mining

Quality control samples are **standard samples** that are periodically run to asses the performance of a mass spectrometry instrument. As such, they provide a potentially useful source of information.
Using QuaMeter[4] quality control data was calculated for **several thousand of mass spectrometry experiments**. Subsequently jqcML was used to store the metrics originating from different sources in a common database to simplify the data management.

Using this data several analyses are possible:

- **Univariate analysis** to evaluate the behavior of each parameter individually. However, this is often insufficient because the different metrics do not function in isolation (Figure 4).
- **Multidimensional analysis** through data mining techniques to evaluate all parameters simultaneously. For example subspace clustering can be used to detect outliers based on a subset of metrics (Figure 5).

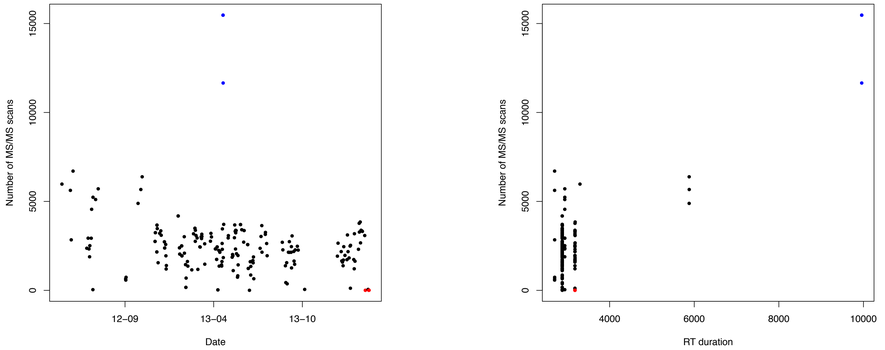

Figure 4: The number of MS/MS scans for a set of standard samples run on a Thermo Scientific LTQ Orbitrap Velos. The samples indicated in red were run when the in-source fragmentation broke down. The samples highlighted in blue might seem outliers as well, but the figure on the right shows that these samples simply ran for a longer time.

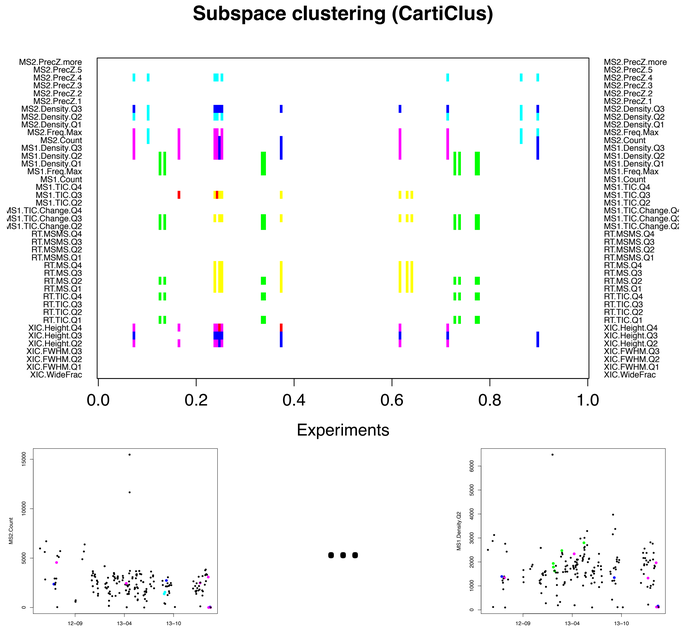

Figure 5: Subspace clustering can be used to detect patterns based on a subset of features. These patterns can subsequently be reevaluated based on the individual metrics to identify their interestingness. The top figure shows each of the specific subspaces across multiple experiments detected by the CartiClus[5] subspace clustering algorithm. The bottom figures show examples of how these subspaces can be traced back to the individual metrics, with the aim of trying to find interesting patterns.

Conclusion

The expressive file format and database structure defined by the qcML specification allows a wide range of possibilities in dealing with quality control data in a standardized way. Furthermore the jqcML library contains all the required functionality in order to work with qcML data. Using these tools we can easily perform data mining techniques on big datasets detailing several hundreds to several thousands of mass spectrometry experiments. Currently we are evaluating different data mining techniques in order to identify interesting patterns, and our future work will continue in this direction.

1. Tabb, D. L. *et al*. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research* **9**, 761–776 (2010).
2. Walzer, M. *et al*. qcML: An exchange format for quality control metrics from mass spectrometry experiments. *Molecular & Cellular Proteomics* (2014).
3. Bittremieux, W. *et al*. jqcML: an open-source Java API for mass spectrometry quality control data in the qcML format. *Journal of Proteome Research* (2014).
4. Ma, Z.-Q. *et al*. QuaMeter: Multivendor performance metrics for LC-MS/MS proteomics instrumentation. *Analytical Chemistry* **84**, 5845–5850 (2012).
5. Aksehirli, E., *et al*. Cartification: A neighborhood preserving transformation for mining high dimensional data. In *ICDM '13*, 937–942 (2013).