MSstatsQC-ML: tree based machine learning methods improve error rates in quality control of mass spectrometry-based proteomics

Eralp Dogu[1], Shantam Gupta[2], Roger Pujol[3,4], Eduard Aguade[3,4],Olga Vitek[2,5]

[1]College of Science, Mugla Sitki Kocman University, Mugla, TR, 2Quantiphi Inc , Boston, MA, [3]Proteomics Unit, Centre de Regulació Genòmica (CRG), Barcelona, SP, [4]Universitat Pompeu Fabra (UPF), Barcelona, SP, [5]College of Science, Northeastern University, Boston, MA.

WP 388

Summary: Automated quality control (QC) systems for mass spectrometry-based proteomics is designed to detect suboptimal performance across multiple metrics (such as intensity, retention time etc.) of multiple analytes. Using univariate per-metric and per-analyte statistical methods is now common across proteomics laboratories. However, the increasing number of samples, metrics and analytes limits the performance of these methods, complicates the interpretation, and inflates the false alarm rates. This work presents a machine learning approach to analyze longitudinal QC performance.

Availability and implementation: The code, the documentation and example datasets are available open-source at www.msstats.org/msstatsqc under the Artistic-2.0 license. The package can be downloaded from www.msstats.org/msstatsqc or from Bioconductor www.bioconductor.org and used in an R command line workflow.

1. MSstatsQC: statistical process control for system suitability monitoring and quality control

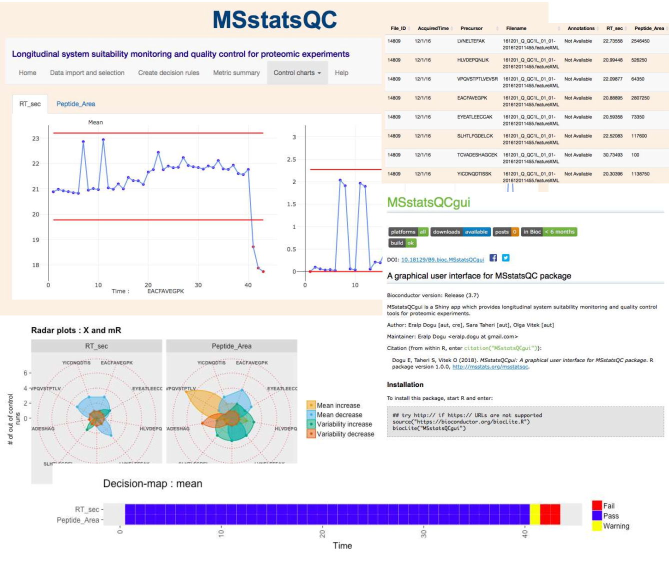
Our framework MSstatsQC (1,2) contributes to the proteomic community additional statistical SPC methods for longitudinal monitoring of both SST and QC with a web based interface.

1) The methods include simultaneous monitoring of mean and variation of a metric (Individual and Moving Range - *XmR*), time weighted control charts to detect small changes (Cumulative Sum - $CUSUM_m$ and $CUSUM_v$ charts), change point analysis for identifying time of a change, and maps for high-dimensional decision making.
2) MSstatsQC also includes a machine learning algorithm to monitor multiple metrics and peptides of interest

3) The methods take as input quantitative metrics such as *retention time, total peak area and peak asymmetry*, or any other quantitative metric of the experimentalist's choice. The methods support experiments with global data-dependent and data independent acquisition as well as targeted experiments.

4) The web interface is available directly from MSstatsQCgui www.bioconductor.org.

1.1 MSstatsQC in action



File_ID	AcquiredTime	Precursor	Filename	Annotations	RT_sec	Peptide_Area
14809	12/1/16	LVNELTEFAK	161201_Q_QC1L_01_01-201612011455.featureXML	Not Available	22.73558	2546450
14809	12/1/16	HLVDEPQNLIK	161201_Q_QC1L_01_01-201612011455.featureXML	Not Available	20.99448	526250
14809	12/1/16	VPQVSTPTLVEVSR	161201_Q_QC1L_01_01-201612011455.featureXML	Not Available	22.09877	64350
14809	12/1/16	EACFAVEGPK	161201_Q_QC1L_01_01-201612011455.featureXML	Not Available	20.88895	2807250
14809	12/1/16	EYEATLEECCAK	161201_Q_QC1L_01_01-201612011455.featureXML	Not Available	20.59358	73350
14809	12/1/16	SLHTLFGDELCK	161201_Q_QC1L_01_01-201612011455.featureXML	Not Available	22.52083	117600
14809	12/1/16	TCVADESHAGCEK	161201_Q_QC1L_01_01-201612011455.featureXML	Not Available	30.73493	100
14809	12/1/16	YICDNQDTISSK	161201_Q_QC1L_01_01-201612011455.featureXML	Not Available	20.30396	1138750

MSstatsQCgui

platforms all | downloads available | posts 0 | in Bioc < 6 months | build ok

DOI: 10.18129/B9.bioc.MSstatsQCgui

A graphical user interface for MSstatsQC package

Bioconductor version: Release (3.7)

MSstatsQCgui is a Shiny app which provides longitudinal system suitability monitoring and quality control tools for proteomic experiments.

Author: Eralp Dogu [aut, cre], Sara Taheri [aut], Olga Vitek [aut]

Maintainer: Eralp Dogu <eralp.dogu at gmail.com>

Citation (from within R, enter citation("MSstatsQCgui")):

Dogu E, Taheri S, Vitek O (2018). *MSstatsQCgui: A graphical user interface for MSstatsQC package*. R package version 1.0.0, http://msstats.org/msstatsqc.

Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("MSstatsQCgui")
```

Figure 1: General MSstatsQC framework for an selection reaction monitoring (SRM) experiment from a QTRAP instrument processed in QCloud (3), required input format, control charts for retention time for EACFAVEGPK, radar plots for overall performance with all the peptides in the mix and decision-map for the longitudinal performance.

References

[1] Dogu, E. et al. (2017). MSstatsQC: Longitudinal system suitability monitoring and quality control for targeted proteomic experiments. *Molecular & Cellular Proteomics, 16*(7), 1335-1347.

[2] Dogu, E. et al. (2019). MSstatsQC 2.0: R/Bioconductor package for statistical quality control of mass spectrometry-based proteomics experiments. *Journal of Proteome Research*, 18 (2), 678-686.

[3] Chiva, C. et al. (2018). QCloud: A cloud-based quality control system for mass-spectrometry-based proteomics laboratories. *PloS one, 13*(1)

[4] Dogu, E. et al. (2019). MSstatsQC-ML: MSstatsQC-ML: A machine learning based method to monitor quality control of mass spectrometry-based proteomics. *In preparation.*

2. MSstatsQC-ML: A machine learning based method to monitor quality control of mass spectrometry-based proteomics

MSstatsQC-ML (4) is an analytical workflow that uses a combination of *statistical sampling, design of experiments, simulation and supervised learning* approaches to detect suboptimal QC performance. Figure 2 illustrates the workflow of MSstatsQC-ML. The algorithm extends the regular use of traditional SPC methods in profiling QC for proteomics and translates QC monitoring problem into a decision making approach per metric per analyte accounting for hypothesis testing.

MSstatsQC-ML creates a general optimal performance profile per peptide using kernel densities after transforming and robust scaling each metric. Next, it generates a series of artificial runs (failing runs) with random artifacts (step changes, variability problems and linear drifts) and another series of runs with optimal performance using factorial experimental design ideas as a new oversampling approach. It also combines real failing runs if they are available. Then, it uses these runs as training and validation sets *for random forest model* which is an ensemble method inspired by decision trees. Using this ML model we next try to classify deviations from optimal performance and hence building an automated system to monitor, detect and control changes in mass spectrometry based proteomics.

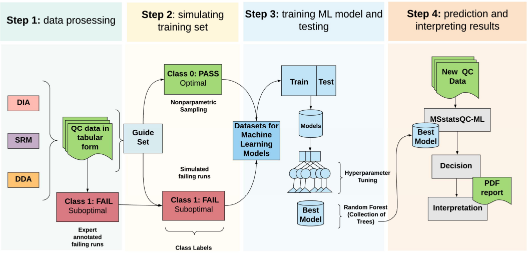


Figure 3: General workflow of MSstatsQC-ML algorithm

2.1 Data

DDA Qcloud dataset: The dataset was generated during DDA using bovine serum albumin (BSA) as a reference protein, quantified with a LTQ-q-Orbitrap Fusion Lumos. Eight peptides were analyzed and each of them was characterized with four metrics (chromatographic peak area, retention time, mass accuracy and full width at half maximum (FWHM)).

Synthetic RT increase dataset: The dataset were designed to evaluate longitudinal QC monitoring algorithms using DDA-QCloud dataset where the true nature of the problem is known. Measurements at 50 time points (runs) were selected randomly and retention time increase was created only for the second half of the dataset for the peptides SLHTLFGDELCK, VPQVSTPTLVEVSR and YICDNQDTISSK

2.2 MSstatsQC-ML outperforms control charts with multiple metrics and/or peptides

MSstatsQC-ML effectively detects deviations from the optimal performance. We used simulated datasets to showcase the performance. As the datasets include 25 runs with suboptimal performance, we see a similar pattern in Figure 3. We provide Hotelling's T2 control chart results as a traditional way of multivariate monitoring where the pattern is not clear. Figure 3 shows the decision map created from MSstatsQC-ML indicating suboptimal performance for the three peptides. To help interpretability, we provided root causes maps which show the longitudinal profile of the input features of the algorithm. This plot shows the values of the statistical features for YICDNQDTISSK indicating an increase in RT.

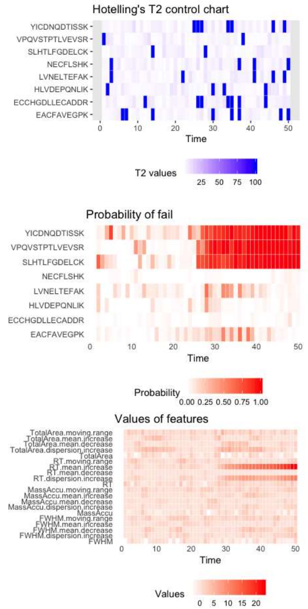


Figure 3: Performance analysis for Synthetic RT increase dataset. First three peptides are unstable and system is unacceptable due to changes RT. Mean level of RT increases over time. Change point is expected to be around 25th run.

3. Use case: longitudinal quality control of real time DDA data

Up to this point, we discussed the proposed method using artificial datasets and preset scenarios. Although artificial datasets are helpful to show the performance of the method, we expand the implementation by adding real applications.

We used the QCloud-DDA-test dataset to show ML based profiling of multiple metrics and multiple peptides. MSstatsQC-ML model was trained with the QCloud-DDA training dataset where a simulation size of 1000 runs were selected. Mean accuracy and area under curve (AUC) values were measured as 0.986 and 0.991, respectively. Then QCloud-DDA test dataset was tested to detect annotated runs. For this particular dataset, it was known that column was changed and the runs after time point 100 was annotated as "LC and or MS troubleshooting". Figure 4 shows the decision map. We observed higher failing probabilities after time 100 while we observed weaker indications of a fail for the rest of the dataset.

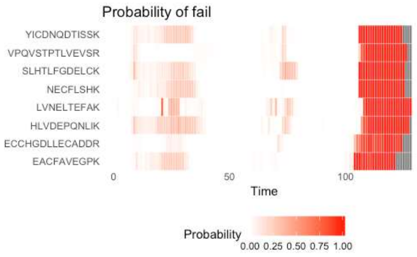


Figure 4: Decision map for QCloud DDA dataset.

3.1 Interpretability

Figure 5 shows root causes maps for this dataset provided additional insight. Root causes maps indicate the values of each statistical feature to interpret sub-optimal performance over time.

When we scan all peptides monitored, we observed the same pattern concluding the problem globally effected the performance. The root causes maps for all peptides shared the same patterns across time for the same statistical features. Examples of the root causes were given in Figure 5 for EACFAVEGPK, LVNELTEFAK, SLHTLFGDELCK and NECFLSHK, respectively. The dark red tiles indicated a subtle increase in the mean of FWHM, decrease in the mean of RT and increase in the mean of mass accuracy. A slightly different case happened for NECFLSHK as for this particular peptide the major effect of the problem appeared as RT decrease.

Overall, the proposed method help us look at the general performance over multiple peptides and metrics with the decision maps, using summaries that can detect diverse types of problems such as subtle shifts. Next, it allowed us to zoom into each peptides and statistical feature to assess the underlying reasons of the problems, and time it potentially happened.

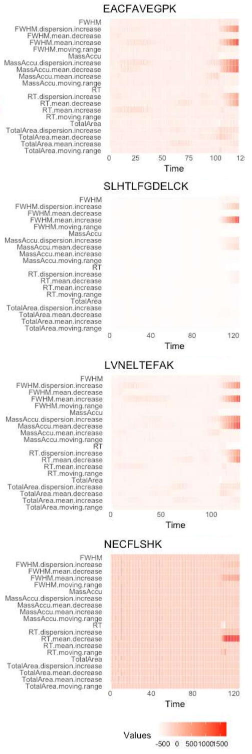


Figure 5: Root causes maps for EACFAVEGPK, LVNELTEFAK, SLHTLFGDELCK and NECFLSHK

Conclusion: In this study, we presented recent developments in MSstatsQC for quality control of mass spectrometry based proteomic experiments. A new functionality based on machine learning called MSstatsQC-ML helps experimentalists use even a handful of optimal runs to monitor performance longitudinally. MSstatsQC-ML can handle DDA, SRM and DIA datasets and any metric of interest. It can automatically summarize large number of peptides and metrics. Our results demonstrated that it is advantageous to monitor multiple metrics of multiple peptides with modern machine learning tools. We proposed visual approaches for exploring the large decision space of multiple peptides.

Communication:
Eralp DOGU eralp.dogu@gmail.com
Olga VITEK o.vitek@neu.edu
www.msstats.org/msstatsqc