Poster G-17

SuperQual: A tool to assess similarity of liquid chromatography mass spectrometry data

Authors:
Lukas Mueller (*IMSB*)
Markus Mueller (*IMSB*)
Reto Ossola (*IMSB*)
Ruedi Aebersold (*IMSB*)

Short Abstract: We have developed a novel software tool, SuperQual, to assess the reproducibility of LC/MS data. The application of SuperQual to a large dataset of LC/MS runs demonstrates how the software enables the quantification of data reproducibility and the detection of poor quality LC/MS runs.

Long Abstract:

Liquid chromatography coupled to mass spectrometry (LC/MS) is the method of choice to analyze complex protein mixtures and to perform comparative studies in large proteomic experiments. LC/MS makes it possible to identify the molecular elements in such complex samples and to quantify their abundance values. Data reproducibility and quality control is therefore a critical factor for downstream quantification of the acquired LC/MS data. Several software tools exist to display and visually inspect LC/MS data in 2 dimensional maps, however, manual inspection is time consuming and often not feasible in high throughput proteomic experiments. Therefore, automated data quality assessment by computational means is required to ensure consistent data quality in high throughput experiments.

SuperQual is a novel software tool programmed in object-oriented C++ and can be downloaded from the project homepage (http://sashimi.sourceforge.net/software_lc-ms.html#SuperQual). The program quantifies the reproducibility of LC/MS data and allows to evaluate data quality and similarity in large LC/MS data sets. Preprocessing of the MS raw data is performed by an upstream data analysis tool and data integration is ensured by XML formatting to enhance the software compatibility to different MS instruments. SuperQual comes with a set of generic C++ classes to interface proteomic data entities such as MS features (m/z, z, TR and VP, MS2 information etc.), LC/MS runs (MS feature handling and processing functionality), and XML parser/writer to archive and extract LC/MS runs in/from XML, which facilitates the integration of the program into other data analysis pipelines. In this study, an inhouse preprocessing routine was used to extract MS features from the raw data, which were stored in XML format and served a data input for SuperQual.

The program routine starts with the import of the preprocessed MS data from XML format. Subsequently, a similarity score is assigned to every possible pair of LC/MS runs by the computational steps of LC/MS alignment and similarity scoring. The first step of LC/MS alignment is necessary to remove variations in the LC elution time of peptide features. SuperQual contains a LC/MS alignment module, which corrects retention time errors. The algorithm searches common peptides between two LC/MS runs within a large mass to charge

and retention time tolerance window and uses then a robust regression method to build a model of the retention time shift. The model enables the prediction of retention time shifts and is used to correct retention time values between the two LC/MS runs.
After retention time normalization in the LC/MS alignment step, the program assigns to each LC/MS pair a similarity score. Initially, common peptides between two LC/SM runs are extracted using the mass to charge, charge state and retention time coordinates at user defined tolerance levels. The common peptides pairs PN are utilized to compute two subscores, which reflect the reproducibility of retention time and peak intensity between the two runs. The computed similarity score is normalized between zero to one where one represents a very good similarity score and zero stands for no LC/MS similarity. Computed similarity scores are stored in a similarity matrix, which is displayed by SuperQual in color matrix using the gnuplot library. Following the similarity assessment of each LC/MS pair, the constructed similarity matrix is transformed by hierarchical clustering into a similarity tree using the Unweighted Pair Group Method with Arithmetic Mean using the similarity scores as intercalate distances. At the end of a SuperQual analysis process, all computed results (number of common features, individual similarity scores, tree structure etc.) are stored in a XML file to facilitate further downstream processing.

SuperQual analysis was performed on 20 LC/MS repeats to quantify data variability introduced by repeated sample preparation and LC/MS analysis. Four samples of N-glycosylated peptides from human serum were prepared and each of which was analyzed five times by LC/MS yielding 20 LC/MS runs. MS features of the acquired LC/MS runs were extracted by the preprocessing routine and subsequent data analysis was performed as described above by SuperQual. The computed similarity matrix gives a good overview of the data quality and good LC/MS similarities are apparent by general high similarity scores colored in red. The color plot facilitates the detection of outliers as was observed for two LC/MS runs, which showed remarkably low similarity scores. Inspection of the base peak chromatograms (BPC) of identified poor quality runs confirmed the findings, where BPC intensity levels little showed little to no signal. This demonstrates the power of automated LC/MS similarity assignment in the detection of outliers, which can be removed (similarity thresholding) or differently processed (weighted by LC/MS similarity score) in a further analysis step. A simplified view of the similarity between the different LC/MS run is available by the constructed similarity tree. The tree forms 4 main branches, which reflect the four different sample preparations where the different LC/MS repeats are grouped into the branch of their corresponding sample.

In summary, SuperQual enabled the quantification of data reproducibility and the detection of low quality LC/MS runs. The presented data demonstrate that even in high reproducible MS systems poor quality data are a common scenario and underline the necessity for automated quality assessment in large LC/MS data sets. Therefore, SuperQual is a useful tool for data quality control and can readily be integrated into other LC/MS analysis software. Beside the assessment of LC/MS reproducibility, the application of SuperQual can also be expanded to the quantification of separation efficiency in different fractionation techniques (ion exchange chromatography, free flow electrophoresis etc.) by looking at LC/MS similarities between different fractions or fractionation steps. In addition, similarity assessment allow to structure LC/MS runs into naturally occurring groups as shown by the hierarchical clustering, which is very valuable information for further data analysis as for example in a multi dimensional LC/MS alignment process.