International Journal of Mass Spectrometry xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

Full Length Article

Semi-supervised quality control method for proteome analyses based on tandem mass spectrometry

Elizaveta M. Solovyeva[a,b], Anna A. Lobas[a,b], Arthur T. Kopylov[c], Mikhail V. Gorshkov[a,b,*]

[a] Moscow Institute of Physics and Technology (State University), Moscow Region, Dolgoprudny, 141701, Russia
[b] V.L.Talrose Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow, 119334, Russia
[c] Institute of Biomedical Chemistry, Moscow, 119121, Russia

ARTICLE INFO

Article history:
Received 30 March 2017
Received in revised form 6 September 2017
Accepted 15 September 2017
Available online xxx

Keywords:
Proteomics
Mass spectrometry
Experiment quality control

ABSTRACT

With the increasing volume of bottom-up proteome analyses using liquid chromatography/mass spectrometry (LC–MS), a rapid and easy for interpretation assessment of the experimental system performance becomes an essential part of the analytical workflow. A variety of the so-called quality control (QC) tools measuring this performance in advance of sophisticated and expensive analyses have been developed and used in practice in many proteomic laboratories worldwide. In the attempts of covering all aspects of LC–MS instrument operation, these tools are typically loaded with a high number of metrics that require performing large-scale analyses using complex standard digest mixtures followed by extensive data processing. Yet, a rapid and simple assessment of the instrument's readiness for a large-scale proteome analysis is often everything needed in routine laboratory practice. In this work we propose a new MS/MS-based quality score which allows performing this assessment without employing full-scale experimental work with annotated samples and time-consuming data analysis. The proposed metric shows high specificity and accurate assessment of the analytical runs and can be used with single protein digest standards. It also allows clear graphical visualization of the quality of the run that makes it useful for day-to-day practice.

1. Introduction

Tandem mass-spectrometry (MS/MS) combined with liquid chromatography (LC) is widely used for proteome analysis. One of the main objectives of deep proteome characterization using high resolution, high throughput LC–MS/MS system is the identification of proteins in the whole cell or tissue lysates. This characterization is commonly performed by so-called "bottom-up" approach, which includes a lot of sample preparation and handling steps, such as lysis, purification, fractionation, enzymatic digestion of the proteins, and peptide separation followed by LC–MS/MS analysis [1–3]. In the course of this analysis the state-of-the-art LC–MS/MS instruments produce large volumes of data. The standard day-to-day laboratory practice includes optimization of the experimental parameters for improving the quality of these data. Because of the large number of parameters affecting the outcome of the analysis, the proper quality control (QC) over the performance of LC–MS/MS system is crucial for saving valuable instrument time and the samples which can be present in a limited amount, such as the ones obtained from patients. Significance of the QC step in the workflow of LC–MS/MS analysis has been grasped by the community for years, and a variety of approaches or metrics to quantify the performance of the LC–MS/MS system or the quality of the spectra have been proposed. In a seminal work by CPTAC consortium (Clinical Proteomic Technology Assessment for Cancer) a set of 46 LC–MS/MS system performance metrics was described and evaluated [4]. The proposed set of metrics allows monitoring the status of all crucial elements and steps of the proteomic analysis workflow including the performance and stability of the chromatographic system, ionization source, mass spectra acquisition, fragmentation of ions, and data analysis. In the follow-up efforts, a number of QC tools and software implementing these metrics in practice were developed [5–11]. Later, it was shown that these metrics allow unambiguous distinguishing between "poor" and "good" results of the proteomic analyses in either supervised [12], or unsupervised [13] manner.

Having a large number of QC metrics for monitoring LC–MS/MS system performance allows comprehending almost all aspects of system operation, data acquisition and analysis. On the other hand, this requires experienced and highly skillful involvement of the operator of a mass spectrometer to correctly and timely interpret

* Corresponding author at: V.L.Talrose Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow, 119334, Russia.
E-mail address: gorshkov@chph.ras.ru (M.V. Gorshkov).

http://dx.doi.org/10.1016/j.ijms.2017.09.008
1387-3806/

the readings from the entire set of numerous metrics. Moreover, the larger the set of metrics, the higher the complexity of the sample needed for the corresponding QC method. This sample complexity may itself bring the uncontrolled perturbations to the output of the QC method. Thus, it is tempting to use standardized, ideally, single protein digest mixtures for the QC experiment and a QC method based on minimal number of metrics, yet, allowing unambiguous conclusions about the performance of the instrument during proteomic analysis. In this work we propose and evaluate a new QC method which can rapidly characterize the performance of LC–MS/MS-based proteome analysis using easy-to-interpret visualization.

2. Materials and methods

2.1. Sample preparation and LC–MS/MS analysis

The QC method developed in this work was focused on monitoring the performance of the high resolution Orbitrap FTMS hybrid mass spectrometers employed typically for proteome analyses. Three types of LC–MS/MS data were generated and used for evaluating the QC method performance. The first data set was obtained for the whole-cell tryptic digests of human embryonic kidney (HEK293) and commercial malignant glioblastoma cell line (DBTRG-05 mg, ATCC). This data set resembles typical "bottom-up" characterization of complex proteomes. The analyses used to generate this data set were performed on the same LC–MS/MS instrumentation under similar experimental settings and conditions over the period of one year. Details for the sample preparation, including the cell lysis, protein digestion, and purification used in this work have been reported elsewhere [14]. The second data set included the analyses of standard Cytochrome c digest (1.6 nmol, Lyophilized, Thermo Scientific Dionex) prepared according to the manufacturer's instruction. One-hour LC gradient was applied for these analyses. High-resolution Q-Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific, Waltham, MA) was used for the analysis of the above samples. The third data set was obtained earlier at the Pacific Northwest National Laboratory (PNNL) and represents a number of LC–MS/MS runs for the Shewanella oneidensis MR-1 lysate digest performed using LTQ-Orbitrap FTMS instrument. These runs were annotated manually by the laboratory's experts as "good", "ok", or "poor" [12].

2.2. Data processing

All raw files were converted to MGF format using MSconvert from the ProteoWizard package (release: 3.0.5533) with "PeakPicking" filter [15]. X!Tandem search engine (v. CYCLON 2012.10.01.1) [16] was used for processing all LC–MS/MS data and peptide identification. The searches were performed against the reference SwissProt human proteome database (version 04/2013). The search parameters for the first and the second data set included maximum two missed cleavages, cysteine carbamidomethylation as fixed modification, and oxidation of methionine and acetylation of the N-terminus of the protein as variable ones. The precursor mass tolerance was set at 10 ppm and the fragment mass tolerance at 0.01 Da. For the third data set, the parameters recommended in the original paper [12] were used. Identifications were filtered to 1% FDR at the peptide spectrum match (PSM) level using MP score software [17].

3. Results and discussion

3.1. Calculation of the metric

"Bottom-up" proteomics includes many sample preparation and handling steps, and errors in each of these steps, as well as the underperformance of the MS instrument may result in the low quality data. It is typically characterized by low number of identified peptides or proteins in case of proteome analyses. Because of the large number of experimental parameters and the added complexity of the state-of-the-art MS instruments, determining the specific reason for the poor outcome of the analysis is challenging. In the environment when high resolution MS instruments are increasingly overloaded with different kinds of proteome analyses, the current trend in quality control (QC) software developments is not only the calculation of numerous metrics of the system performance, but the quality score generation for the current data [12,13]. This score may timely characterize the overall performance of the instrument and quickly guide a researcher in the decision making for continuing the analysis instead of going through all the metric outcomes. Furthermore, such a score is more suitable to be understood by the instrument's operators and technicians without specific knowledge of the mass-spectrometry-based proteomic technologies. In case of bottom-up proteome analysis, MS/MS spectrum is one of the obvious QC metrics for rapid assessment and scoring of the instrument performance. Indeed, the quality of MS/MS spectra directly affects the success in correct peptide identification and, thus, the depth of the whole proteome coverage. Seven MS/MS-based spectral metrics were proposed by Rudnick et al. [4] including median ion injection time, number of peaks, S/N of identified peaks, and a number of identified MS/MS spectra for four intensity quartiles. These metrics are very useful to comprehensively characterize the quality of the analysis, but they lack simple graphical representation and all have to be measured to assess the performance of the instrument.

We believe that two characteristics of MS/MS spectrum, which are crucial for the outcome of successful peptide identification, may alone provide a comprehensive view on the spectrum quality. They are the total number of peaks in the spectrum and their average intensity, which can be obtained from MGF file. Using these two metrics the quality of the whole LC–MS/MS run can be easily visualized as a two-dimensional plot. Importantly, measuring these characteristics of the spectrum does not require peak identification in the mass spectra, further simplifying the implementation of the QC method. Fig. 1b shows an example of such a plot for LC–MS/MS analysis of HEK293 sample. Each point in Fig. 1b corresponds to an MS/MS spectrum identified at 1% FDR. It can be seen that most of the tandem spectra obtained in this run fall within a certain angle (indicated by the blue lines) in this two-dimensional space formed by the proposed characteristics. The coordinates of the vertex of the angle were calculated as first percentiles for distributions of the number of peaks and intensity, respectively, as shown in Fig. 1a. Then, the MS/MS scans corresponding to the first and the last percentiles for both distributions were excluded from the calculation of the quality angle. The slopes of the lines are adjusted to have one percent of the considered identified MS/MS spectra under the bottom and above the upper rays, respectively (Fig. 1a). Thereafter, one defines for each MS/MS scan, whether it is falling within the angle, or not, according to the following equation:

$$K_1(X - X_1) + Y_1 < Y(X) < K_2(X - X_1) + Y_1 \quad (1)$$

in which X_1 and Y_1 are the coordinates of the angle vertex, K_1 and K_2 are the adjusted slopes of the rays, X and Y are the number of peaks and average peak intensity in the MS/MS scan, respectively. On average, about 4% of all identified spectra fall outside the angle while this fraction exceeded 75% for the unmatched MS/MS spectra. The distributions of search engine score (e-value) for scans falling inside and outside of the above-described angle, shown in Fig. S1 (Supporting information), illustrate that the outliers have in general significantly higher scores than the spectra lying inside the angle. Note that the lower the e-value of identified PSM, the higher the confidence of data. Finally, the quality score for the particular

Please cite this article in press as: E.M. Solovyeva, et al., Semi-supervised quality control method for proteome analyses based on tandem mass spectrometry, Int. J. Mass Spectrom. (2017), http://dx.doi.org/10.1016/j.ijms.2017.09.008