Research on Efficient Algorithm of Peptide Identification and Quantification in Proteomics

Zhen WANG[1,2,*], Ji-yang ZHANG[2], Yun-ping ZHU[1] and Hong-wei XIE[2]

[1] Institute of Radiation Medicine State Key laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences, Beijing 102206, China

[2] Department of Automatic Control, College of Mechatronics and Automation, National University of Defense Technology, Changsha 410073, China

*wang.zhen.zhen2008@163.com

Keywords: Proteomics, Label-free quantification, XICFinder, Detectability, Reproducibility, Precision.

Abstract. Because of its simple experiment design and low experiment cost, label-free quantitative technology based on mass spectrometry analysis is being used more and more widely. Aiming at peptide identification and quantification, which are pivotal step of the label-free quantitative analysis, we develop an efficient algorithm named XIC Finder based on C++ platform. In term of the peptides which failed to be identified by MS/MS spectrum, we utilize ESP model trained by 20 optimized peptide features with Random Forest method, to predict those peptides detectability and chose the highest scored peptide as the correct peptide. Compared with MaxQuant and IDEAL-Q, other algorithms developed for quantitative MS data, XIC Finder improves the performance of the peptide identification and quantification significantly. Furthermore, we evaluated the reproducibility and precision of XIC Finder by using the replication dataset and the UPS1 standard data set respectively and prove the result is better than other algorithms.

Introduction

The Human Genome Project (HGP) [1], which called the program of "Apollo Moon Landing" in life science research, was declared complete in 2003. Life science has gone into the post-genome era, and the focus has transferred from genetic information discovery to functional analysis [2]. Comprehensive proteomics study will become an important field of life science in the twenty-first century. Proteome attempts to character the expression of the full set of proteins for a biological sample under defined conditions qualitatively and quantitatively. Because of its high throughput, sensitivity and resolution for analytes, liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) [3] is widely used in quantitation proteomics. With the development of mass spectrometry recently, such as Orbitrap, a large number of high-precision data are produced and increase the complexity of quantitative analysis [4].

Quantitation techniques in proteomics are divided into relative quantitation and absolute quantitation [5, 6], and relative quantitation which can find differentially expressed proteins is widely used in clinical medicine. Discovery proteomics, also named Shotgun is the classic experiment strategy in relative quantitative proteomics [7], which contains four main steps: 1) Sample extraction and Digestion. Sample proteins are digested by trypsin and converted by proteolysis into peptides. 2) Analysis by LC-MS/MS. Peptides analyzed by LC-MS/MS produce profile spectrum, then the most abundance peptides detected in MS1 scan are selected for fragmentation. 3) Database search. The fragment ion spectra are assigned to their corresponding peptide sequences by database searching software, such as Mascot and X!Tandem [8, 9, 10]. 4) Peptides and proteins quantitation. Based on the former results, peptides and proteins quantitation are processed with relevant mathematical and statistical analysis methods.

Peptide identification and quantitation is essential procedures in quantitation proteomics. Through searching database of theoretic spectrum, peptides selected for fragmentation can be identified, and the rest peptides can be searched by accuracy mass and time tag (AMT) library [11] for identification in this article. In peptide quantification step, the performance of the algorithm which estimates peptide abundance by the area under the curve of extracted ion chromatogram is proved feasible. Thus, ion chromatogram extraction in MS1 scan is fundamental procedure for peptide quantification [12].

Data and Method

Datasets

Five kinds of samples (including samples A-E) were analyzed in the LTQ-Orbitrap mass spectrometery, a sample of each experiment was repeated three times technologically. Samples consist of yeast proteins and 48 UPS1 standard proteins, where the amount of 60ng/μL USP1 protein in the sample A-E were 0.24, 0.74, 2.2, 6.7 and 20 fmoles respectively [13]. Such protein samples can simulate the actual sample of the protein under different conditions of this difference in expression. The data set can be downloaded freely on Proteome Commons server, and the Hash number is 'zccpb0hwGbgThe8AQOCJbOFqUir7erYp0w49UVqqmx1jlQM8xUEjzX99chJMfzYPlpUkSmjAqPsTVVoo+YAAAAAAAAgZQ=='.

Through Trans-Proteomic Pipeline v4.4, RAW files is converted MGF files [14], which can be searched by database searching software. The setting parameters used by Mascot Demon v2.2 to perform database searching were as follows: 1) Mass tolerance of precursor and fragment ions is 10ppm and 0.5Da respectively. 2) Complete digestion with trypsin. 3) Fixed modification is Carbamidomethyl of amino acid C, and variable modification is oxidation of amino acid M.

Peptide Feature Extraction and Prediction on Peptide Detectability

Physical and chemical properties of peptides are important factors influencing peptide detectability [15]. However the extreme complexity of peptide properties, including the mass and length, isoelectric point, hydrophobicity and hydrophilicity of peptides, consists of 550 feature vector [16]. After correlation analysis, we found that the physical and chemical characteristics of peptides contain numerous highly correlated features—these redundant information will increase complexity and difficulty in process of data analysis. We use Principal Component Analysis (PCA) for feature extraction, which can greatly reduce the complexity of data analysis. PCA algorithm uses Schmidt orthogonal transformation, which can convert the relevant variables to the linearly independent variable [17].

Based on properties of peptides, we establish predicting model of peptide detectability named ESP via Random Forest method, which is one of the efficient Machine Learning methods [18]. The larger number of decision trees are, the higher accuracy of prediction is, but the time of model training will be longer as well. If you use 50000 decision trees, the model construction takes about 12 hours. Using R language pack of Random Forest algorithm, we select the 500 decision trees to ensure sufficient accuracy and make the training time reduced. We select the peptide data set" TRAIN- TEST_ yeast_ CPTAC_ HighLowND_ z-1_0_ AVG_AADB2_110607.csv" as the training set which has 3153 identified peptides, including 550 kinds of physical and chemical properties.

The Kernel of XICFinder Algorithm

In discovery proteomics, the mass spectrometry instrument is operated on data-dependent-acquisition (DDA) mode [19], where several precursor ions detected in a MS1 scan are selected for fragmentation. DDA technology is limited by its stochastic acquisition, however, it doesn't mean that the rest peptides in the MS1 scan can't be identified, but the low MS2 sampling rate cause numerous peptides failed to be identified.

In principle, XICFinder finds all the possible extracted ion chromatograms (XICs) from the MS spectrums firstly. By means of database searching and peptide detection predicting, XICFinder carry through peptide identification and quantification respectively, where peptides are represented by XICs and the area of XIC is considered as the corresponding peptide quantitation value. XICFinder consists of six modules, and utilizes three main C++ Class to process MS data analysis. Class IdxRaw can read the Raw file and detect XICs initially trough invoking class MSInf, which stores relevant information of MS spectrums. XICFinder invokes API of Xcalibur to read MS information, and use the similarity of the m/z in the mass tolerance (defaulted 50ppm) and isotope matching to find the same peptide among MS. Then, we extracted intensity of the same peptides based on RT and combined them into a XIC. Class MResult is designed for peptide identification, which invokes API of MSparse to read the result of Mascot searching contained identified information of peptides selected for fragmentation. The rest peptides corresponding to XICs but not identified by MS/MS spectrums are compared with theoretical tryptic peptides, and develop lists for the candidate peptides. Then, we utilize the ESP model that can predict peptides detection to classify the candidate peptides, after multiple matches, we chose the highest scored peptide as the correct peptide in correspondence to the XIC.

Results and Discussion

Peptide Feature Extraction Based on Principal Component Analysis

Based on the R language platform, using principal component analysis, physical and chemical properties of 550 peptides were extracted and optimized. We calculate the linear combination of the original variables, which become new linearly independent components pc1, pc2, pc3 ⋯ shown in Figure 1:

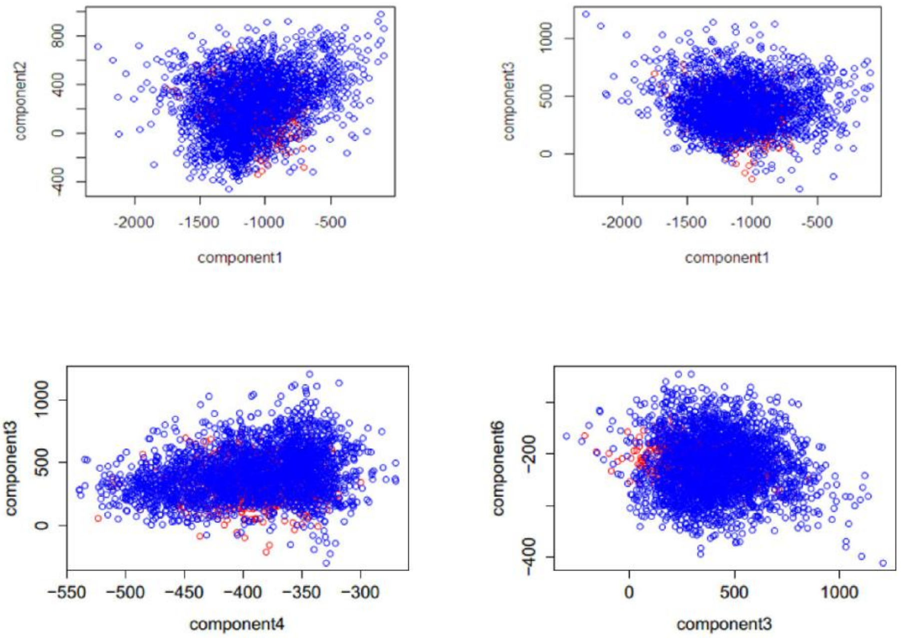

Figure 1. Scatterplots showing linear independence among the main component

Then, we selected the top twenty principal components as the factor, which can reduce 550 features to 20 factors and achieve 99% explanation. The relative results are shown in Figure 2.

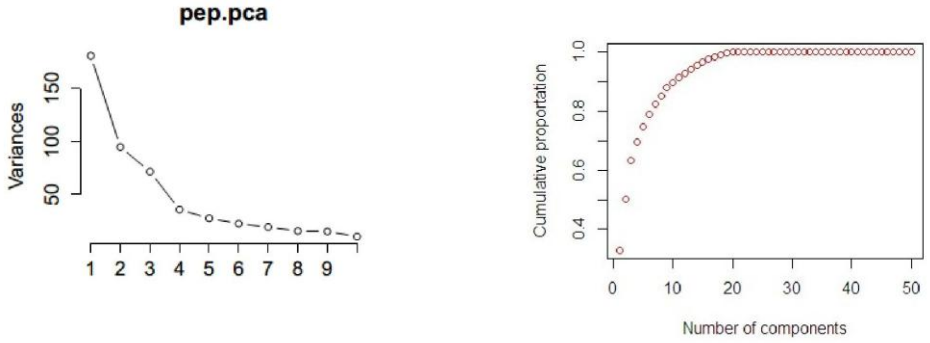

Figure 2. Variance analysis of the main component (left) and cumulative explanation of number of components (right)

Peptide Detectability Predicting Based on ESP Model

In consideration of peptide that failed to be identified by MS/MS spectrum, XICFinder generate a list of candidate peptides corresponding to theory trypsin peptides via searching Accuracy Mass and Time tag (AMT) library. Furthermore, we apply the established ESP model to predicting detectability of those candidate peptides.

For instance, from 15.489min to 15.634min, XICFinder constructed a complete XIC(charge +1, m/z 340.945, S/N 41.877) in Retention Time, which was not identified by Mascot database searching but by Accuracy Mass and Time tag (AMT) library searching. After multiple matching, we obtain a list of candidate peptides, where we chose the highest scored peptide by ESP model as the correct peptide corresponding to the XIC.

Results Analysis of Peptide Identification and Quantitation by XICFinder

XICFinder performs well on peptide identification and quantitation respectively. In terms of the number of peptides identified only by XICs, the peptides identified by XICFinder are significantly more than peptides, which were identified by MaxQuant [20] and IDEAL-Q [21].The number of peptide identification by three algorithms is shown in Figure 3(left). XICFinder found 3798 XICs, and identified 2744 peptides—1470 peptides were identified by MS/MS spectrum and 974 were identified by AMT matches based on peptide detectability prediction. The identification rate of peptides is up to 72.4%, more than MaxQuant (61.5%) and IDEAL-Q (65.9%). However, ROC curve of XICFinder is slightly worse than MaxQuant due to its multiple AMT matches based on peptides detectability—this suggests that False Positive Rate (FPR) is higher than MaxQuant slightly, which is shown in Figure 3(right), and that will be the emphasis of the algorithm to advance in the future.

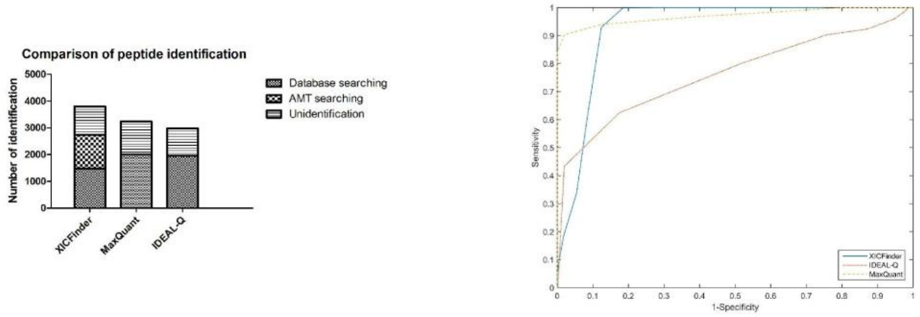

Figure 3. Comparison of peptide identification (left) and ROC curve (right) among three algorithms.

In consideration of peptide quantification, we calculate the coefficient of variation of log2 peptide abundance in 10 repeated experiments, and draw a box plot for further analysis as Figure 4. The mid-values of CV evaluated by three algorithms are 0.0147, 0.0195 and 0.0173 respectively, which indicates the reproducibility of XICFinder is better than others.In view of the corresponding peptides of UPS1 Proteins, we need to utilize unary linear regression to analyze the linearity between five abundance estimates and actual sample loads. The linearity can be measured by Coefficient of Determination R-square of linear regression. The higher coefficient of determination is, the better the accuracy of quantitative result is. The mid-values of CD evaluated by three algorithms are 0.9843, 0.9806 and 0.9787 respectively, which indicates the accuracy of XICFinder is better than other algorithms. The relative results are shown in Figure 4.

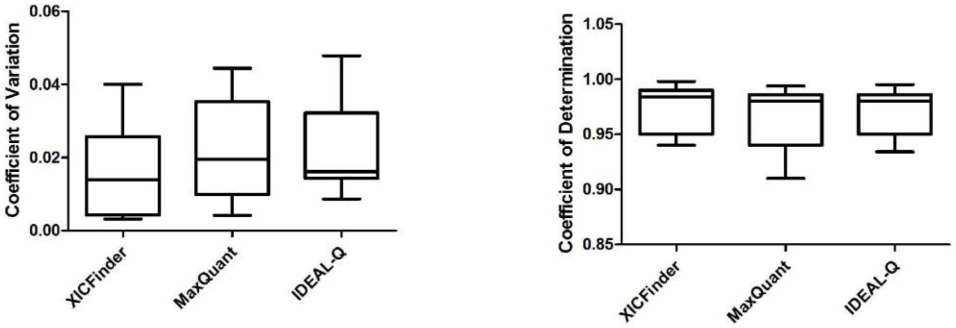

Figure 4. Coefficient of Variation (left) and Coefficient of Determination (right) in peptide quantification

Conclusions

Mass spectrometry analysis has become the core method for quantitative proteomics research, where peptide identification and quantitation is essential procedures. We develop relevant algorithm named XICFinder, which can identify about 72.4% peptides in ten repeated experiment. Furthermore, the higher coefficient of variation and coefficient of determination, which were calculated by identified peptides and UPS1 standard peptides respectively, indicates that the reproducibility and accuracy of XICFinder are better than MaxQuant and IDEAL-Q.

The low identification rate of peptides is due to the low sampling rate resulted from data-dependent-acquisition (DDA) of the mass spectrometry instrument. Thus, data–independent-acquisition (DIA) has become the dominant technique gradually, which can make large scale MS2 quantitation available [22]. In DIA mode, peptides detected in MS1 scan are all fragmented into MS/MS spectrums, which can be identified by database searching and quantify more peptides than DDA mode.

Acknowledgement

The authors sincerely acknowledge Jiyang Zhang for his unselfish assistance in improvement of XICFinder algorithm. This work was supported by National Natural Science Foundation of China. Grant Numbers: 31171266, 31000587, 31000591.

References

[1] Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome [J]. Nature, 2001, 409(6822): 860-921.

[2] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome[J]. Nature, 2012, 486(7402): 207-214.

[3] Ong S E, Mann M. Mass spectrometry-based proteomics turns quantitative [J].Nat Chem Biol, 2005, 1(5): 252-262.

[4] Olsen J V, de Godoy L M F, Li G, et al. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap[J]. Molecular & Cellular Proteomics, 2005, 4(12): 2010-2021.

[5] Elliott M H, Smith D S, Parker C E, et al. Current trends in quantitative proteomics [J]. J Mass Spectrom, 2009, 44(12): 1637-1660.

[6] Vaudel M, Sickmann A, Martens L. Peptide and protein quantification: a map of the minefield [J]. Proteomics, 2010, 10(4): 650-670.

[7] Tabb D L, McDonald W H, Yates J R. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics[J]. Journal of proteome research, 2002, 1(1): 21-26.

[8] Eng J K, Searle B C, Clauser K R, et al. A face in the crowd: recognizing peptides through database search [J]. Mol Cell Proteomics, 2011, 10(11): R111 009522.

[9] Sadeh N M, Hildum D W, Kjenstad D, et al. Mascot: an agent-based architecture for coordinated mixed-initiative supply chain planning and scheduling[C]//In Workshop on

Agent-Based Decision Support in Managing the Internet-Enabled Supply-Chain, at Agents' 99. 1999.

[10] Brosch M, Swamy S, Hubbard T, et al. Comparison of Mascot and X! Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold[J]. Molecular & cellular proteomics, 2008, 7(5): 962-970.

[11] Zimmer J S D, Monroe M E, Qian W J, et al. Advances in proteomics data analysis and display using an accurate mass and time tag approach[J]. Mass spectrometry reviews, 2006, 25(3): 450-482.

[12] Wong J W H, Sullivan M J, Cagney G. Computational methods for the comparative quantification of proteins in label-free LCn-MS experiments[J]. Briefings in bioinformatics, 2008, 9(2): 156-165.

[13] Tabb D L, Vega-Montoto L, Rudnick P A, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography−tandem mass spectrometry[J]. Journal of proteome research, 2009, 9(2): 761-776.

[14] Zhang W, Zhang J, Xu C, et al. LFQuant: A label‐free fast quantitative analysis tool for high‐resolution LC‐MS/MS proteomics data[J]. Proteomics, 2012, 12(23-24): 3475-3484.

[15] Liu K, Zhang J, Wang J, et al. Relationship between sample loading amount and peptide identification and its effects on quantitative proteomics[J]. Analytical chemistry, 2009, 81(4): 1307-1314.

[16] Zeng X C, Wang S X, Zhu Y, et al. Identification and functional characterization of novel scorpion venom peptides with no disulfide bridge from Buthus martensii Karsch[J]. Peptides, 2004, 25(2): 143-150.

[17] Jolliffe I. Principal component analysis[M]. John Wiley & Sons, Ltd, 2002.

[18] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling[J]. Journal of chemical information and computer sciences, 2003, 43(6): 1947-1958.

[19] Schwudke D, Liebisch G, Herzog R, et al. Shotgun Lipidomics by Tandem Mass Spectrometry under Data‐Dependent Acquisition Control[J]. Methods in enzymology, 2007, 433: 175-191.

[20] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification[J]. Nature biotechnology, 2008, 26(12): 1367-1372.

[21] Tsou C C, Tsai C F, Tsui Y H, et al. IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation[J]. Molecular & Cellular Proteomics, 2010, 9(1): 131-144.

[22] Gillet L C, Navarro P, Tate S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis[J]. Molecular & Cellular Proteomics, 2012, 11(6): O111. 016717.