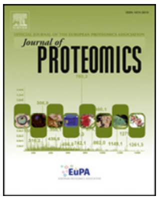
Contents lists available at ScienceDirect

Journal of Proteomics

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



A multi-model statistical approach for proteomic spectral count quantitation

Owen E. Branson, Michael A. Freitas *

[a] The Ohio State Biochemistry Graduate Program, The Ohio State University, Columbus, OH, USA
[b] Department of Molecular Virology, Immunology and Medical Genetics, The Ohio State University, Columbus, OH, USA
[c] Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA

ARTICLE INFO

Article history:
Received 11 January 2016
Received in revised form 23 May 2016
Accepted 24 May 2016
Available online 31 May 2016

Keywords:
Spectral Counting
Proteomics
edgeR
DESeq
baySeq
q-Value

ABSTRACT

The rapid development of mass spectrometry (MS) technologies has solidified shotgun proteomics as the most powerful analytical platform for large-scale proteome interrogation. The ability to map and determine differential expression profiles of the entire proteome is the ultimate goal of shotgun proteomics. Label-free quantitation has proven to be a valid approach for discovery shotgun proteomics, especially when sample is limited. Label-free spectral count quantitation is an approach analogous to RNA sequencing whereby count data is used to determine differential expression. Here we show that statistical approaches developed to evaluate differential expression in RNA sequencing experiments can be applied to detect differential protein expression in label-free discovery proteomics. This approach, termed MultiSpec, utilizes open-source statistical platforms; namely edgeR, DESeq and baySeq, to statistically select protein candidates for further investigation. Furthermore, to remove bias associated with a single statistical approach a single ranked list of differentially expressed proteins is assembled by comparing edgeR and DESeq q-values directly with the false discovery rate (FDR) calculated by baySeq. This statistical approach is then extended when applied to spectral count data derived from multiple proteomic pipelines. The individual statistical results from multiple proteomic pipelines are integrated and cross-validated by means of collapsing protein groups.

Biological significance: Spectral count data from shotgun proteomics experiments is semi-quantitative and semi-random, yet a robust way to estimate protein concentration. Tag-count approaches are routinely used to analyze RNA sequencing data sets. This approach, termed MultiSpec, utilizes multiple tag-count based statistical tests to determine differential protein expression from spectral counts. The statistical results from these tag-count approaches are combined in order to reach a final MultiSpec q-value to re-rank protein candidates. This re-ranking procedure is completed to remove bias associated with a single approach in order to better understand the true proteomic differences driving the biology in question. The MultiSpec approach can be extended to multiple proteomic pipelines. In such an instance, MultiSpec statistical results are integrated by collapsing protein groups across proteomic pipelines to provide a single ranked list of differentially expressed proteins. This integration mechanism is seamlessly integrated with the statistical analysis and provides the means to cross-validate protein inferences from multiple proteomic pipelines.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Mass spectrometry based proteomics is the most diverse platform for protein identification and characterization, in part due to advances in protein isolation techniques, chromatographic separation options and a diverse array of ionization, fragmentation and data acquisition techniques [1]. In shotgun proteomics mere protein identification is usually not sufficient to understand the complexity of biological phenomena. Label-free spectral counting is a robust semi-quantitative technique directly applicable and widely used in shotgun proteomics [2–7]. Spectral count data from shotgun proteomics experiments are heavily influenced by chromatographic separations and sample complexity, as well as the choice of analytical instrumentation and the implementation of dynamic exclusion parameters, and therefore should be considered semi-quantitative and semi-random [8–11]. Regardless of bias associated with collection of mass spectrometry peptide data, ultimately protein identification must be inferred from the generated peptide spectra with the use of database search engines [12–19]. The next

Abbreviations: PSM(s), Peptide Spectrum Match(es); Pfu, *Pyrococcus furiosus*; UPS, Universal Proteome Standard Mix; PRIDE, PRoteomics IDEntifications data repository; EAE, Experimental Autoimmune Encephalomyelitis; FDR, False Discovery Rate; Xcorr, Cross-Correlation Scoring Analysis; UQ, Upper Quartile; TMM, Trimmed Mean of M-values; RLE, Relative Log Expression; FWER, Family-Wise Error Rate.

* Corresponding author at: The Ohio State University Medical Center, 460 West 12th Avenue, Columbus, OH 43210, USA.

E-mail address: freitas.5@osu.edu (M.A. Freitas).

http://dx.doi.org/10.1016/j.jprot.2016.05.032
1874-3919/© 2016 Elsevier B.V. All rights reserved.

challenge is determining differential protein expression between cohorts of complex proteomes and prioritizing these protein candidates for validation.

In discovery based proteomic experiments the number of samples collected is often small (<10). In these instances it is not possible to prove that the counts fit a Gaussian (normal) distribution. Powerful statistical alternatives have been routinely applied in the proteomics community to perform differential expression analysis of spectral count data [20–25]. Models based on the Poisson distribution have historically been applied to model count data. The main limitation to the Poisson distribution is that it has only one model parameter and cannot effectively model under- or over-dispersed data. When there is not sufficient data to confirm that the sample variances are equal, quasi-likelihood or generalized linear mixed effects modeling approaches can be used [20, 22]. Another alternative to the Poisson distribution is the negative binomial distribution. The approaches described herein leverage the negative binomial distribution to model over-dispersed count data through determining the unique mean-variance relationship [26–31]. Very similar to spectral count data from shotgun proteomics, RNA sequencing data is over-dispersed, multivariate in nature, and often limited by few biological replicates. This has encouraged the development of so-called tag-count based statistical approaches to determine differential expression. These tag-count based statistical approaches are powerful alternatives to traditional parametric and non-parametric tests when analyzing RNA sequencing data [32–36].

The approach described here, termed MultiSpec, employs a multi-model tag-count based statistical approach. Individual statistical results are combined and re-ranked using a median q-value/FDR approach. This holistic representation of differential expression can be extended to the analysis of spectral count data obtained from multiple proteomic pipelines. MultiSpec is built upon open-source statistical platforms, namely edgeR, DESeq and baySeq and is executable in the R programming language (v 3.0) [37]. The highlighted analysis (EAE/Sham) is a product of the multi-model statistical analyses of spectral count results derived from three proteomic pipelines (MassMatrix, MyriMatch and Proteome Discoverer). The three independent statistical analyses and integration of the results across proteomic pipelines utilized a maximum of 242 MB of real memory and was complete in 317 s. Files containing the raw spectral counts from each proteomic pipeline, detailed descriptions of the figures generated by MultiSpec and corresponding result tables are available in Supplemental Material 12. The authors anticipate continuous advancement of this modular R script. Therefore, the most current version is available from the authors upon request. The version of the R script used in this manuscript is available in Supplemental Material 13.

2. Materials and methods

2.1. Experimental data sets

Two publicly available and previously described datasets were used to illustrate the utility of label-free spectral counting and highlight the statistical capabilities of MultiSpec.

First the dataset described by Chen et al. was used to validate that spectral counts generated by each proteomic pipeline is an acceptable approach to estimate fold changes [38]. In addition, this dataset was also utilized to evaluate the potential influence of TMM normalization on estimating fold changes. This dataset consisted of 36 human proteins spiked into a *Pyrococcus furiosus* (Pfu) lysate. Each cassette consisted of six human proteins from the Universal Proteome Standard from Sigma Aldrich (UPS). Each cassette was spiked into a Pfu lysate at different ratios: 1:1, 4:3, 3:5, 2:1, 4:1 and 1:8 as described in Supplemental Table 1. For the purpose of this study, these spike ratios were assumed to be correct. Five technical replicates were produced for each condition (UPSA or UPSB) using a 95-min gradient and spectra collected with an LTQ-Orbitrap Velos. This dataset consisting of 10 RAW files was analyzed by three separate label-free proteomic pipelines (MassMatrix, MyriMatch and Proteome Discoverer) as described below. Data were searched against a custom FASTA database containing 2152 forward protein sequences: 36 Universal Proteome Standard (UPS) protein sequences, 71 common contaminant proteins and the Pfu UniProt database (08/20/2012) [38]. This custom database was concatenated to a reverse decoy database to estimate peptide and protein false discovery rates (FDR). The common constraints applied to database searches included: (1) limiting the search to b/y ions, (2) in-silico sequence digestion after Lys and Arg except those proceeding a Pro, (3) fixed modification due to carbamidomethylation of Cys (+57.0215 Da), (4) variable modifications for the formation of Glu to pyro-Glu (−18.011) and for oxidation of Met (+15.9949) and (5) precursor mass and fragment mass tolerances were set at 10 ppm and 0.6 Da, respectively. The full set of search parameters is provided in Supplemental Table 2.

Second, to highlight the ability for MultiSpec to identify unknown proteomic changes a dataset from a murine model of multiple sclerosis (EAE/Sham) was obtained from the PRoteomics IDEntifications (PRIDE) data repository [39]. The EAE/Sham dataset consisted of 18 RAW files from six biological replicates (three EAE and three Sham surgery) each analyzed in technical triplicate. The overall workflow of the data analysis is outlined in Fig. 2. The EAE/Sham data were analyzed by three separate label-free proteomic pipelines (MassMatrix, MyriMatch and Proteome Discoverer), against the complete, reviewed, forward/reverse UniProt murine database (May 2014), containing 16,677 forward sequences. Like the Pfu dataset, the search parameters were harmonized across search engines (Supplemental Table 2).

2.2. Label-free proteomic pipelines

Leveraging multiple search engines has been shown to increase peptide and protein identifications [40–43]. In this approach three proteomic pipelines (MassMatrix, MyriMatch and Proteome Discoverer) each consisting of unique peptide spectrum match (PSM) filtering criteria, search engine and protein grouping mechanism were used to cross validate differential protein expression. The complete set of parameters for each proteomic pipeline is outlined Supplemental Table 2.

2.2.1. MassMatrix [12,44–46]

RAW data was converted to mzXML data format using MSConvert in ProteoWizard (v 3.0.7494) [47,48]. In the case of the EAE/Sham dataset, the mzXML files for the three technical replicates were merged to represent a single biological replicate. An MS spectra was mapped to not more than one peptide sequence. Peptides with a *p*-value less than 0.05 were retained and mapped to either forward or reverse protein sequences. Decoy and non-decoy protein identifications and their associated spectral counts were parsed and recombined using an in-house Python script as described below (Supplemental Material 3). In cases where a homologous protein group was identified, it was represented by a comma separated list of unique identifiers (UniProtIDs). The maximum protein score of the proteins in a homologous protein group was used to represent the protein group. The final harmonized protein list was ranked by protein score and filtered where each valid protein ID contained at least two unique peptides. The FDR was estimated using a target-decoy strategy and all proteins were retained until the incorporation of protein decoy exceeded the 5% protein FDR [49]. To account for the missing value problem across multiple samples, the homologous protein groups were split across search results based on their protein ranking and grouping in the database search. The regrouping used a bipartite approach similar to that described by Zhang et al. but at the protein group level rather than the peptide level [16]. A table of the database search results and protein groupings for each sample was supplied as an input. The final output was a combined spreadsheet of spectral counts with a harmonized grouping of proteins across all samples. If

Download English Version:

https://daneshyari.com/en/article/1225121

Download Persian Version:

https://daneshyari.com/article/1225121

Daneshyari.com