Published in final edited form as:

J Proteome Res. 2019 February 01; 18(2): 748–752. doi:10.1021/acs.jproteome.8b00483.

EBprotV2: A Perseus Plugin for Differential Protein Abundance Analysis of Labeling-based Quantitative Proteomics Data

Hiromi W.L. Koh^{1,2}, Yunbin Zhang³, Christine Vogel^{3,*}, and Hyungwon Choi^{1,2,4,*}

¹Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

²Saw Swee Hock School of Public Health, National University of Singapore, Singapore

³Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, USA

⁴Institute of Molecular and Cell Biology, Agency for Science, Technology, and Research, Singapore, Singapore

Abstract

We present EBprotV2, a Perseus plug-in for peptide ratio-based differential protein abundance analysis in labeling-based proteomics experiments. The original version of EBprot models the distribution of log-transformed peptide-level ratios as a Gaussian mixture of differentially abundant proteins and non-differentially abundant proteins and computes the probability score of differential abundance for each protein based on the reproducible magnitude of peptide ratios. However, the fully parametric model can be inflexible and its R implementation is time consuming for datasets containing a large number of peptides (e.g. >100,000). The new tool built in C++ language is not only faster in computation time, but also equipped with a flexible semi-parametric model that handles skewed ratio distribution better. We also developed a Perseus plug-in for EBprotV2 for easy access to the tool. In addition, the tool now offers a new submodule (MakeGrpData) to transform label-free peptide intensity data into peptide ratio data for group comparisons and performs differential abundance analysis using the mixture modeling. This approach is especially useful when the label-free data has many missing peptide intensity data points.

Graphical Abstract

*To whom correspondence should be addressed: cvogel@nyu.edu; hwchoi@nus.edu.sg.

Availability: The software is freely available at <https://github.com/cssblab/EBprot/> (Apache 2.0 license), along with a tutorial and example data sets.

Keywords

differential abundance; labeling-based proteomics; statistics; peptide-level analysis; semi-parametric modeling; label-free data; Perseus plugin; R package

Introduction

Labeling-based proteomics is a powerful method for detection of differentially abundant (DA) proteins. Compared to label-free approaches, stable isotope labeling has the advantage that isotopically labeled peptides from two or more samples are detected together in a single mass spectrometry (MS) experiment, and their abundance ratios provide relative quantitation immediately. There are various types of labeling-based quantification including isobaric labeling such as iTRAQ¹ and TMT², and non-isobaric labeling such as SILAC³ and dimethyl labeling⁴, and these are appealing choices for many biological applications.

Labeling-based proteomic data often consist of peptide ratios computed between a sample of interest against a reference sample. Currently available data analysis pipelines for ratio data rely on heuristic methods for assigning statistical significance values, such as the one built within the Perseus environment⁵⁻⁶, which operates on protein-level ratios instead of peptide-level ratios. In MS-based proteomics experiments, however, the measurements are made for peptides, not proteins, and some proteins are quantified with more peptides than others. Further, the ratios may vary wildly across peptides within a single protein, especially in low abundance proteins. This cross-peptide reproducibility information is typically ignored in the calculation of confidence scores for differential abundance, partly because protein ratios can be easily derived from their respective peptide-level ratios (e.g. by averaging) and it is more intuitive to work with a single ratio value for each protein than multiple ratios *per* protein.

This observation previously motivated us to develop a probabilistic framework called EBprot.⁷ EBprot models peptide ratio data, not protein ratio data, to reward proteins with reproducibly high or low ratios across multiple peptides. At the same time, proteins represented by a small number of ratio values (e.g. ratio from a single peptide) are

considered differentially abundant only if the magnitude of ratio(s) is sufficiently large in either direction, depending on the overall ratio distribution in the given data set. By contrast, proteins with consistently large ratios across multiple peptides can be considered differentially abundant even with modestly large or small ratio values on average. Therefore, EBprot's scoring system naturally balances the peptide coverage and the variability of ratios across different proteins in its calculation of differential abundance scores accordingly.

Here we report an extension of EBprot, called EBprotV2. The new implementation has the following improvements: (i) the software package was re-written in C++ language, substantially reducing computational time for high dimensional datasets; (ii) log-transformed ratios are modeled using a semi-parametric mixture model, making it more flexible and robust in handling skewed data than Gaussian mixtures; (iii) a new module to transform label-free data to ratio data at the peptide level, most beneficial for group comparisons in the presence of a large number of missing peptide intensity values; and (iv) a new graphical user interface (GUI) in the Perseus environment to broaden the access to the software.

Figure 1A shows the workflow of EBprotV2. The user may start with a label-free peptide intensity data (top left), or a peptide ratio data (top middle). In the former, the data transformation module called EBprot.MakeGrpData converts label-free peptide intensity data into peptide ratio data between groups. The ratio data is then analyzed by the mixture model-based assessment of differential abundance for proteins, and the final report is generated in a text file along with other related files useful for plotting purposes. The details of the Perseus Plugin can be found in the GitHub repository (<https://github.com/cssblab/EBprot/>).

EBprotV2 analyzes large data sets faster than EBprot

We first compared the computation time of EBprot (R package) and EBprotV2 (command line) on a standard Windows-operated desktop computer equipped with Intel i5-4670 processors (3.80GHz). We used two data sets of different size, previously analyzed in the paper describing EBprot⁷: a UPS1 human protein standard mixture data set consisting of 1960 proteins and 8550 peptides, and a time course iTRAQ data set for phosphoproteomics dynamics in EGF-stimulated HeLa cells consisting of 19,616 (replicate 1) and 24,392 (replicate 2) phosphopeptides. In the latter, there are three time points (0 min, 10min and 24h) where comparisons were made with respect to the baseline (i.e. 0 min). The total number of comparisons is twice the number of phosphopeptides. This number corresponds to 20~24 times the number of proteins in the UPS data set, and thus the computation time is expected to be longer. In the UPS data set, both implementations completed analysis in about 8 seconds. In the iTRAQ data set, EBprotV2 was able to finish the analysis substantially faster than EBprot: it took 127 seconds for the former, while it took 1,176 seconds for the latter.

Semi-parametric mixture model is more adaptive to non-Gaussian mixture distributions

We next tested the flexibility of the semi-parametric mixture model using synthetic data sets, where we are able to control the parameters determining the shape of log-transformed peptide ratios (logarithm base 2). We created synthetic data sets with 2,000 proteins, each with a varying number of peptides from a mixture of four Poisson distributions (see Figure S1A). The differentially abundant (DA) proteins were all assumed to be up-regulated and the proportion of DA proteins was varied at 10%, 15% and 20%. Ratios were simulated using a mixture of two Gaussian distributions of mean 1.5 and 2.5 with standard deviation of 0.3 to form a bimodal distribution. The rest of the data were simulated with a Gaussian distribution of mean 0 and variance 0.5.

Figure S1A shows the distribution of log₂ peptide ratio data for 20% differentially abundant proteins in the synthetic data. The model fit by the two implementations, shown in Figure S1B, clearly demonstrates the benefit of the flexibility provided by the semi-parametric model. When we compared the accuracy of differential abundance calls at various probability score thresholds, the receiver operating characteristic (ROC) curves show that, when the distribution is poorly fit (EBprot), the sensitivity and specificity can be considerably influenced (Figure S1C). Interestingly, we observed that the poor model fit by EBprot affects the performance more seriously when the proportion of differentially abundant proteins was larger, i.e. when the one of the two modes closer to the null distribution is completely missed. Although this likely represents an extreme case in real experiments, the semi-parametric model in EBprotV2 provides the safeguard against a poor model fit while providing comparable sensitivity and specificity at all score thresholds.

Meanwhile, the semiparametric model also provides better specificity than the Gaussian mixture. Figure S2 shows the results when we simulated the data with no differentially abundant proteins with varying noise levels. The plots illustrate that, as the noise level increases, the chance of falsely identifying differentially abundant proteins due to noisy peptide ratio values increases in the fully parametric model of EBprot, whereas EBprotV2 estimated 0 probability score for every protein across all three levels of noise. In sum, the semi-parametric model provides comparable or better performance to EBprot's implementation, while providing a robust model fit in all extreme cases that can occur in real experimental data sets.

Analyzing label-free data with many missing peptide intensities as group-to-group ratio data

Figure 1B illustrates the process of transforming label-free intensity data into ratio data for group comparisons. This module derives peptide ratios between comparison groups from label-free data as specified by the user (see Software Manual distributed in the GitHub page). This data transformation is particularly useful when peptides are not reproducibly detected and quantified across samples, rendering standard hypothesis testing-based differential abundance analysis difficult. While it is possible to remedy this situation by

deriving protein-level intensities, those values derived from raw peptide intensity data plagued by many missing values are likely to be poor representation of true underlying quantities.

In the derivation of group-specific mean peptide abundance (to be used for ratio calculation), the module first assigns a weight to each sample that is proportional to the number of quantified peptides in the same protein in the sample. Here, the sample weight for a peptide reflects the peptide coverage of its parent protein in a given sample, and therefore samples with a higher peptide coverage in the parent protein contribute more to the average intensity of a given peptide. Using these weights, a weighted average intensity is calculated for the given peptide for each comparison group. Finally, peptide-level ratios are computed for each pair of comparison groups (specified by the user) using the weighted average intensity values.

Following the ratio calculation, user can choose to remove outlying peptide ratio values from each protein. In a protein, peptide ratios lying outside 4 times the **reference standard deviation** (learned from data) from the median ratio in the protein are removed from the analysis. The reference standard deviation is learned from proteins with at least 5 peptides in the data set, and the description can be found in Koh *et al*⁷. The group-to-group peptide ratio data generated above, or a peptide ratio data originally acquired from a labeling-based experiment, is then analyzed by the semi-parametric mixture modeling implemented in EBprotV2, to produce protein-level differential abundance scores. Figure 1C illustrates the scenario from a simulated data, where the non-parametric component of the model (EBprotV2) can fit the ratio distribution of the peptides from differentially abundant proteins (red line) more adaptively than the fully parametric model in EBprot. The skewed distributions may not arise in many experimental data, for which EBprot's Gaussian mixture modeling works well. The semi-parametric alternative provides a safeguard against bad model fit in case the ratio distribution is skewed or even multi-modal (see below).

We applied the tool to a superSILAC data set with 40 breast cancer tumor tissue samples of three major subtypes (14 ER+/PR+, 15 Her2+, and 11 triple negatives (TN))⁸. The original study aimed to identify proteins uniquely abundant in the TN subtype. Note that the original data is from labeling-based superSILAC experiments, yet the data is formatted as if it is label-free data due to the nature of superSILAC method. Using the data transformation module (EBprotV2.GrpComparisons), we constructed the peptide ratio data comparing ER+/PR+ with TN and Her2+ with TN, using the peptides quantified in at least 4 samples in all three groups ($k=4$ in Figure 1B). We also compared the results with the analysis performed at the protein-level, similar to our work in the original EBprot. For the protein-level analysis, protein ratios were derived by taking the median peptide ratios in each protein. For the peptide-level analysis, an outlier removal step was applied to remove aberrant peptide ratios.

Using the score threshold associated with 5% false discovery rate, the peptide-level analysis of EBprotV2 identified 86 and 191 proteins (out of 10,600 proteins) to be differentially abundant in TNs compared to ER+/PR+ and Her2+ subtypes, respectively. As expected, the significant proteins were different in terms of peptide coverage in both group comparisons. The proteins selected in the peptide-level analysis usually had multiple peptides showing

reproducibly high or low ratios, yielding more credibility to the selected differentially abundant proteins (Figure 2A). Furthermore, some previously proposed positive markers of TN subtype, such as Annexin 1 (ANXA1)⁹, was selected only in the peptide-level analysis, but not in the protein-level analysis.

At the same error rate, the protein-level analysis (using EBprotV2) identified 151 and 285 differentially abundant proteins from the same group comparisons. However, the majority of the proteins selected in the protein-level analysis, but not selected in the peptide-level analysis, were those with ratio values from a single peptide or those with very few and conflicting peptide ratios. See Table S1 for detailed comparison of the two analysis results.

As shown in Figure 2B, the proteins can be categorized into three groups: differential in both comparisons between TN and others, differential between TN and Her2+ only, and differential between TN and ER+/PR+ only. Strong markers of TN from our results include ELP6, AGR2, CRIP1 (negative) and LEPREL1, FOLR1 and SARG (positive). We also observed previously characterized TN-specific positive markers: ENO1, ANXA1 and CYP1A1; negative markers: CMBL, INPP4B, EEF1A2, FOXA1 and PDXDC1. Among these, however, only the peptide-level analysis selected PDXDC1, FOXA1, INPP4B, ANXA1 and ENO1.

Discussion

In this work, we improved the EBprot software for peptide ratio-based differential protein abundance analysis in various ways. The revised architecture of EBprotV2 is computationally faster and more robust to aberrant distribution shapes of ratios. The Perseus plugin renders the method significantly more accessible by experimental scientists.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by Singapore Ministry of Education grant MOE2016 T2-1-001 (HC) and NIH grant R35 GM127089 (CV).

References

1. Ross PL; Huang YN; Marchese JN; Williamson B; Parker K; Hattan S; Khainovski N; Pillai S; Dey S; Daniels S; Purkayastha S; Juhasz P; Martin S; Bartlett-Jones M; He F; Jacobson A; Pappin DJ, Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 2004, 3 (12), 1154–69. [PubMed: 15385600]
2. Thompson A; Schafer J; Kuhn K; Kienle S; Schwarz J; Schmidt G; Neumann T; Johnstone R; Mohammed AK; Hamon C, Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* 2003, 75 (8), 1895–904. [PubMed: 12713048]
3. Ong SE; Blagoev B; Kratchmarova I; Kristensen DB; Steen H; Pandey A; Mann M, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 2002, 1 (5), 376–86. [PubMed: 12118079]

4. Boersema PJ; Raijmakers R; Lemeer S; Mohammed S; Heck AJ, Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc* 2009, 4 (4), 484–94. [PubMed: 19300442]
5. Tyanova S; Temu T; Sinitcyn P; Carlson A; Hein MY; Geiger T; Mann M; Cox J, The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 2016, 13 (9), 731–40. [PubMed: 27348712]
6. Cox J; Mann M, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008, 26 (12), 1367–72. [PubMed: 19029910]
7. Koh HW; Swa HL; Fermin D; Ler SG; Gunaratne J; Choi H, EBprot: Statistical analysis of labeling-based quantitative proteomics data. *Proteomics* 2015, 15 (15), 2580–91. [PubMed: 25913743]
8. Tyanova S; Albrechtsen R; Kronqvist P; Cox J; Mann M; Geiger T, Proteomic maps of breast cancer subtypes. *Nat Commun* 2016, 7, 10259. [PubMed: 26725330]
9. de Graauw M; van Miltenburg MH; Schmidt MK; Pont C; Lalai R; Kartopawiro J; Pardali E; Le Devedec SE; Smit VT; van der Wal A; Van't Veer LJ; Cleton-Jansen AM; ten Dijke P; van de Water B, Annexin A1 regulates TGF-beta signaling and promotes metastasis formation of basal-like breast cancer cells. *Proc Natl Acad Sci U S A* 2010, 107 (14), 6340–5. [PubMed: 20308542]

Figure 1.

(A) The workflow in EBprotV2. The user can start the analysis from two different places: with and without data transformation step. If the data consists of peptide intensities (or any other data formatted as label-free data), the user converts the data into peptide ratio data using the EBprot.MakeGrpData module. Once the peptide ratio data is ready, the model fitting and score calculation modules can be run to produce protein-level differential analysis results. (B) Data transformation module in two steps: derivation of group-specific weighted average peptide intensity and group-to-group peptide ratio calculation. (C) Semi-parametric mixture model (EBprotV2) versus Gaussian mixture model (EBprot). The example data shown is a simulated data.

Figure 2.

Breast cancer super-SILAC data. **(A)** Difference in the significant proteins between EBprotV2 analyses using peptide-level ratios and protein-level ratios (conventional approach). Proteins selected in both analyses are not shown and only the first 30 up- and 30 down-regulated proteins that are significant only in the protein-level analysis are shown. Red dots indicate the peptides removed from the respective proteins by the outlier removal step. Inclusion of those peptides would have reduced the significance score and thus the proteins would have been labeled as non-significant. **(B)** EBprotV2 analysis using peptide-level ratios for the two comparisons of interest in search of TN-specific protein markers. **(C)** Examples of six proteins selected in the differential abundance analysis. The box plots are the peptide-level weighted average intensities (logarithm with base 2, mean centered) in the three groups.