Improved normalization and discovery in label-free LC-MS proteomics

Tom Taverner,[1] Alan R Dabney[2], Sam O Purvine[1], Gordon A Anderson[1], Mary S Lipton[1], Richard D Smith[1]
[1] Pacific Northwest National Laboratory, Richland, WA 99352 [2] Texas A&M University College Station, TX 77843

Overview

- **Many sources of systematic and random error contribute to the variability observed across proteomics experiments.**
- **Incorporation of factors such as charge state, LC elution time, and mass-to-charge ratio into a normalization scheme significantly reduces technical replicate variability and improves dataset comparability, and provide more powerful statistical assessments.**
- **Different peptide quantitation methods are compared using test datasets.**
- **Combining separate statistical analyses improves performance for discovering differentially expressed proteins.**

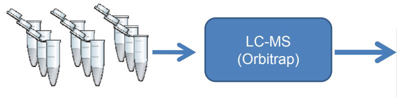
Introduction

- Several methods are widely used for quantifying peptides and proteins [1,2]:
  - *Peptide counting model:* the number of observed peptides for each protein, either as MS/MS spectra (spectral counting) or as spectral features (peptide counting) [3]
  - *Peptide averaging model*: Average all peptide intensities of a protein within a run; results can be analyzed using a one-factor ANOVA or t-test
  - *Peptide-level ANOVA model* [4-10]: the peptide log-intensity is taken to be a linear sum of peptide and protein-level group effects
- We investigated sources of bias and instrument effects that affect LC-MS proteomics intensity measurements by applying simple statistical methods of model building, stepwise regression, and model checking to go beyond the usual data correction steps in quantitative label-free LC-MS proteomics analysis.
- When these sources of bias and variance were corrected, the within-replicate variability was substantially reduced by a factor of 2 or better. This allows for improved detection of differentially expressed proteins using the simple and powerful peptide-level fixed-effects ANOVA scheme, which can either be used to improve quantitative accuracy or reduce the number of technical replicates that need to be performed.

Methods

***Salmonella* datasets**: 19 identical replicates of *S. typhimurium* proteomic digest run in 3 separate LC-MS replicate blocs

Preliminary model checking and model fitting: determining which factors varied between technical replicate blocs on nominally identical datasets and were important for normalization



LC-MS (Orbitrap)

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = \mu_i + \beta_{p[i],\,Tr[j]} + \varepsilon_{ij}$$
(ANOVA model)

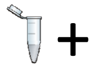
$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = \mu_i + \beta_{p[i],\,Tr[j]} + \varepsilon_{ij}$$
(ANOVA model with elution time)

$$y_{q,i,j} = \mu_{q,i} + \alpha_j + \beta_{Tr[p[i],\,k[j]]} + g_j(\theta_i) + \varepsilon_{-q,i,j}$$
(ANOVA model with elution time, charge, *m/z*)

Terms in the ANOVA model include: the log-intensity of the i'th peptide in the j'th dataset, y_{ij}, is some mean value μ_{ij} added to some random error ε_{ij}; μ_i is the mean value of the *i*th peptide; α_j is the systematic bias of the *j*th experiment; the treatment effect for the protein containing the ith peptide, p[i], under the Tr'th treatment group which the jth experiment falls into, Tr[j], is $\beta_{p[i], Tr[j]}$; and the normalization task corresponds to removing an experiment-specific additive bias from y_{ij}, α_j.

***Shewanella* datasets**: 39 replicates of a *S. oneidensis* proteomic digest spiked with a standard mixture of 10 proteins over a dynamic range of 10^4; run in 3 separate LC-MS replicate blocs

Validation of normalization and discovery on a well-characterized dataset

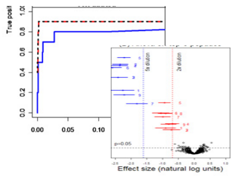


Shewanella digest
(0.05 mg/mL)

Standard proteins
bovine serum albumin
bovine carbonic anhydrase
bovine beta-lactoglobulin
bovine serotransferrin
rabbit glyceraldehyde-3-phosphate dehydrogenase
E. coli beta-galactosidase
bovine alpha-lactalbumin
equine skeletal muscle myoglobin
chicken ovalbumin
bovine cytochrome *c*
rabbit phosphorylase *b*

Dilution	x10^{-1}	x10^{-2}	x10^{-3}	x10^{-4}
1 fold	3x	3x	3x	3x
2 fold	3x	3x	3x	3x
5 fold	3x	3x	3x	3x

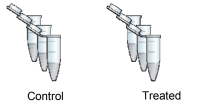
LC-MS (Orbitrap) followed by normalization using our scheme and peptide-based ANOVA



Validation of normalization and discovery by direct measurement of false discovery and true discovery rates for differentially spiked-in proteins

***Caulobacter* datasets**: 3 replicates each of *C. crescentus* proteome grown under normal or carbon-starved conditions

Check performance of proteomic discovery methods on a biological dataset



Control
Treated
(glucose starvation)
C. crescentus cultures

LC-MS (Orbitrap) followed by normalization using our scheme and peptide-based ANOVA/peptide feature counting

Using a widespread label-free LC-MS proteomic analysis averaging over peptide intensities (Rrollup) 17/1097 proteins "discovered" as differentially expressed compared to 181/1097 using peptide-level ANOVA.

Normalization using additional predictors of peptide charge state and LC-MS elution time increased by 18% the number of proteins identified by peptide-level ANOVA as differentially expressed at p-value < 0.05 (181 with the additional predictors compared to 153 without).

CONTACT: **Tom Taverner, Ph.D.**
Biological Sciences Division, K8-98
Pacific Northwest National Laboratory
P.O. Box 999, Richland, WA 99352
E-mail: Thomas.Taverner@pnl.gov

Results

Using elution time in normalization improves data correlation by at least 2-fold

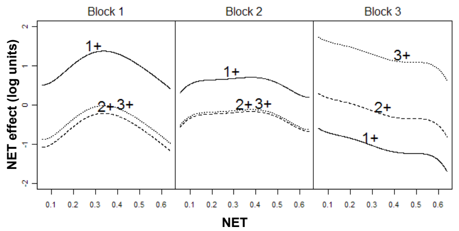


Figure 1. Normalized elution time (NET) effect on peptide intensity by LC-MS block. For the *Salmonella* dataset, the effect of the peptide elution time on the correction for peptide intensity, treating peptide charge states separately, over 3 randomly selected datasets from separate LC column blocks (left to right) for the *Salmonella* dataset.

Charge state 1+ peptides are solid lines, 2+ are dashed lines, 3+ are dotted lines. The relationships between intensity and elution time looked more similar within LC-MS blocks than across blocks, although there were still some systematic differences between samples from the same block.

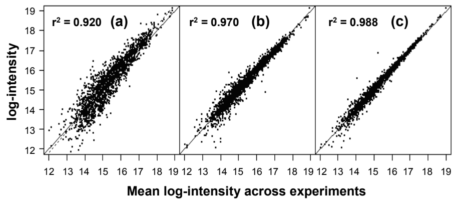


Figure 2. The effect of incorporating more predictors on the agreement between replicate *Salmonella* LC-MS datasets. **(a)** Uncorrected *Salmonella* data from one dataset plotted against mean log peptide intensities (over 19 datasets). **(b)** Same dataset with robust linear regression on elution time and charge state predictors. **(c)** *Salmonella* data corrected for charge, NET and *m/z* biases, fitted to the random-effects model for log peptide intensities incorporating peptide-block interactions.

The correlation coefficient for the fit is inset, y=x is shown as a solid line. The correlation coefficient for the fit is inset and y=x is shown as a solid line.

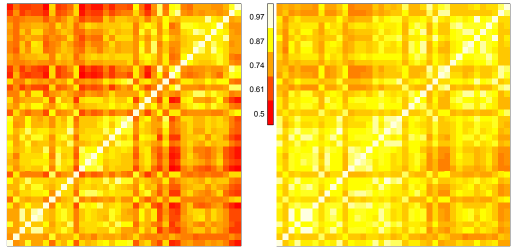


Figure 3. Correlation plots over 39 replicate spiked *Shewanella* datasets showing the effect of incorporating elution time as an additional normalization predictor (**b**) compared to not incorporating the predictor (**a**).

The color scale is kept the same between correlation plots although the minimum correlation value changes from 0.50 in A to 0.69 in B.

Peptide charge state distribution varies

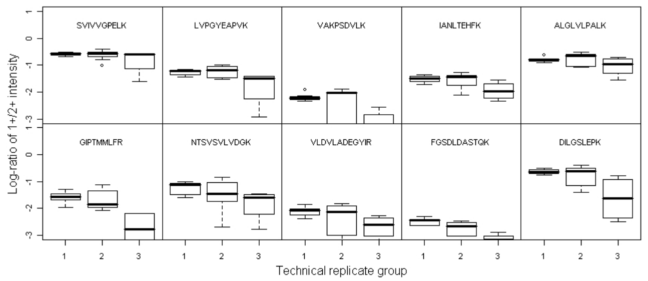


Figure 4. Box plots of log(1+/2+) peptide intensities for ten sampled peptides from the *Salmonella* dataset. The numbers on the x-axis indicate technical replicate block numbers; peptides were observed in 8 replicates of block 1, 7 replicates of block 2, and 4 replicates of block 3. The range of each box plot is inset. The charge state ratio distribution differs between replicate blocks, with more 2+ ions in technical replicate block 3.

Correcting peptide intensity predictors reduces bias

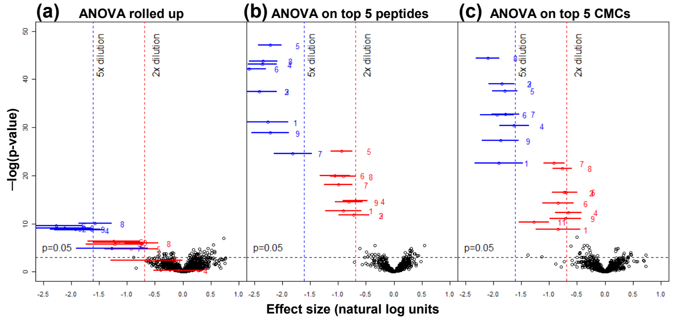


Figure 5. Volcano plots (log fold-change effect vs. negative log p-value) for three different types of statistical treatments of the spiked-in *Shewanella* proteome dataset after normalization. The sample contained 0.05 mg/mL spiked-in proteins; red points indicate 2x dilution and blue points, 5x dilution. The 95% standard error bars from the statistical analysis are provided in corresponding colors. Matrix *Shewanella* proteins are indicated by black points.

In **(a)**, a typical protein rollup analysis was performed where a robust median value of the peptide intensities were taken as an indicator of overall protein intensity, after which ANOVA was performed to compute differences. In **(b)**, a similar ANOVA analysis was performed on the top-5 peptides by median intensity for each protein, fitting to a linear model. In **(c)**, we separated each peptide for each protein by the observed charge state (CMCs) and performed a similar ANOVA analysis on the top 5 median-intensity CMC observations for each protein.

- Peptide-level ANOVA has more statistical power to detect differentially expressed proteins than rolling up peptides to proteins.
- Separating out charge states improves the quantitative accuracy for the 5-fold dilution experiment, bringing 9 out of 10 spiked in proteins to within the 95% confidence interval (CI) of the true value.

Combining hits from peptide count and peptide-level ANOVA improves discovery

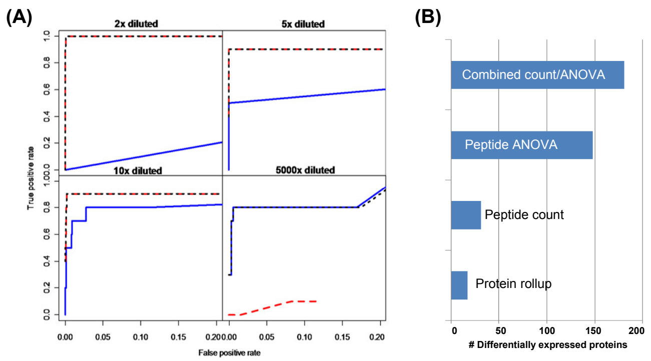


Figure 6. (A) Receiver operating characteristic curves comparing the ability to "discover" as differentially expressed,10 spiked-in proteins out of a constant *Shewanella* proteome background using three quantitative methods: top-peptide ANOVA (red dashed line), a statistical count test (blue solid line) or combining significant hits from both (black dashed line). The plots compare 0.05 mg/mL spiked-in protein mix to 2x, 5x, 10x and 5000x dilutions (the last corresponding to virtually no peptides detected). The combination approach ROC curve is coincident with the top-peptide ANOVA in the 2x, 5x and 10x plots and with the peptide count test in the 5000x plot. **(B)** Comparison of different quantification significance detection methods in determining differentially expressed proteins between control and carbon starved *Caulobacter*.

Conclusions

- **We have identified four practically significant parameters – mass, charge state, elution time, and a random effect due to LC-MS column-peptide interactions – that are not normally taken into account in label-free LC-MS proteomics normalization.**
- **A correction scheme incorporating some of these parameters reduces the variability between LC-MS datasets, improving the statistical power and sensitivity.**
- **These results should be applicable to other kinds of quantitative label-free proteomics experiments.**
- **Missing peptide data is still a challenge for peptide-level ANOVA. On the other hand, peptide counting methods can deal with this problem, but ignore measured intensities altogether.**
- **Our compromise of combining two separate statistical analyses outperforms either ANOVA or peptide count statistical methods for differentially expressed protein discovery.**

Acknowledgements

This research was supported by the U.S. Department of Energy Biological and Environmental Research (DOE/BER). Samples were analyzed using capabilities developed under the support of the NIH National Center for Research Resources (RR18522) and DOE/BER. Significant portions of the work were performed in the Environmental Molecular Science Laboratory, a DOE/BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is operated for the DOE by Battelle under contract DE-AC05-76RLO-1830.

The authors would like to thank Drs. Leticia Britos and Lucy Shapiro at Stanford University providing the *Caulobacter* samples.

References

1. Listgarten J and Emili A. *Mol Cell Proteomics* **2005**, *4*, 419.
2. Vitek O. *PLoS Comput Biol* **2009**, *5*, e1000366.
3. Pang JX, et al. *J Proteome Res* **2002**, *1*, 161.
4. Clough T, et al. *J Proteome Res* **2009**, *8*, 5275.
5. Oberg AL and Vitek O. *J Proteome Res* **2009**, *8*, 2144.
6. Daly DS, et al. *J Proteome Res* **2008**, *7*, 1209.
7. Oberg, AL, et al. *J Proteome Res* **2008**, *7*, 225.
8. Bukhman YV, et al. *J Bioinform Comput Biol* **2008**, *6*, 107.
9. Karpievitch YV, et al. *Bioinformatics* **2009**, *25*, 2028.
10. Karpievitch YV, et al. *Bioinformatics* **2009**, *25*, 2573.
11. Polpitiya AD, et al. *Bioinformatics* **2008**, *24*, 1556.
12. Petritis K, et al. Anal Chem **2006**, *78*, 14, 5026.
13. Lipton MS, et al., *Methods Biochem Anal* **2006**, *49*, 113.