BMC Bioinformatics

MEETING ABSTRACT Open Access

Statistical elimination of spectral features with large between-run variation enhances quantitative protein-level conclusions in experiments with data-independent spectral acquisition

Lin-Yang Cheng[1]*, Yansheng Liu[2], Ching-Yun Chang[1], Hannes Röst[2], Ruedi Aebersold[2,3], Olga Vitek[4]

From Tenth International Society for Computational Biology (ISCB) Student Council Symposium 2014
Boston, MA, USA. 11 July 2014

Background

Many proteomic investigations summarize the quantitative information across multiple spectral features into protein-level conclusions. Data-independent spectral acquisition (DIA) now generates a lot of interest, as it allows us to quantify many spectral features in a single run. However, the disadvantage of DIA experiments as compared, e.g., to Selected Reaction Monitoring (SRM) is that the features are subject to interferences and noise. We argue that between-run variation provides an additional insight for distinguishing good-quality and noisy DIA features. To appropriately use the quantitative between-run variation, it is important to account for the properties experimental design, and distinguish random artifacts from the biological changes. We have previously proposed a method (Chang et al., ASMS 2013) that accounts for the experimental design to eliminate features with low information content.

Results

In this project we furthermore emphasized that conducting regularization helps us avoid exploring every subset of features exhaustively, and allows us to conduct hypothesis tests later on so that we would be able to control the false discovery rate of the feature selection process. We evaluated our proposed approach by using three datasets that have some notion of ground truth: an extensive simulation study, a controlled mixture where proteins were spiked into a complex background in known concentrations, and a study of 232 plasma samples, where 18 proteins were quantified in both SWAH and SRM mode in presence of heavy labeled reference peptides. We worked on [1] protein-level estimates of fold changes between conditions, [2] sensitivity and specificity of detecting changes in protein abundance, and [3] accuracy of relative quantification of protein abundance in individual biological samples. A family of linear mixed models similar to that in MSstats http://www.msstats.org were fit to all the datasets. Then we conducted the regularization and hypothesis test to control the selection false discovery rate.

Conclusion

The results demonstrated that our proposed feature selection approach enhanced sensitivity and specificity of the conclusions, was robust to the amount of noisy fragments, and increased the correlation of subject quantification between SRM and DIA workflows. Importantly, the performance exceeded that of the frequently used ‘top 3’ approach, which consists of using three spectral features with the highest average intensity between runs. Furthermore, we showed that our proposed approach outperforms using correlation to select the information features.

Authors’ details
[1]Department of Statistics, Purdue University, West Lafayette IN, USA.
[2]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, 8093 Zurich, Switzerland. [3]Faculty of Science, University of Zurich, 8057 Zurich, Switzerland. [4]Department of Computer Science, Purdue University, West Lafayette IN, USA.

Published: 28 January 2015

* Correspondence: cheng68@purdue.edu
[1]Department of Statistics, Purdue University, West Lafayette IN, USA
Full list of author information is available at the end of the article

References

1. Clough T, Thaminy S, Ragg S, Aebersold R, Vitek O: **Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs".** *BMC Bioinformatics* 2012, **13**:S16.
2. Chang CY, Picotti P, Hüttenhain R, Heinzelmann-Schwarz V, Jovanovic M, Aebersold R, Vitek O: **Protein significance analysis in selected reaction monitoring (SRM) measurements.** *Molecular and Cellular Proteomics* 2012, **11, Article M111.014662.**
3. Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, Vitek O: **MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments.** *Bioinformatics* 2014.
4. Lockhart R, Taylor J, Tibshirani R, Tibshirani R: **A significance test for the lasso.** *The Annals of Statistics* 2014, **42.**

doi:10.1186/1471-2105-16-S2-A4
Cite this article as: Cheng *et al.*: **Statistical elimination of spectral features with large between-run variation enhances quantitative protein-level conclusions in experiments with data-independent spectral acquisition.** *BMC Bioinformatics* 2015 **16**(Suppl 2):A4.