Statistical approach to absolute protein quantification

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sarah Gerster & Peter Bühlmann

Seminar für Statistik
ETHZ
gerster@stat.math.ethz.ch

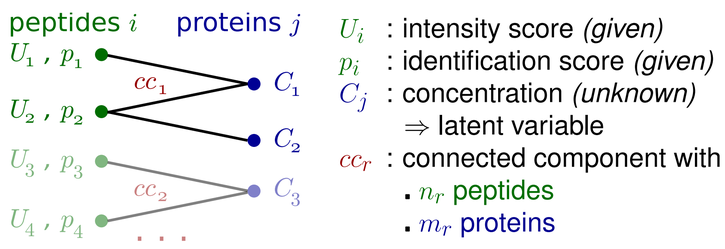
SfS
Seminar for Statistics

Abstract

We propose a statistical approach to label-free protein quantification with three main advantages compared to existing methods. (i) Peptide intensities are modeled as random quantities, allowing to account for the uncertainty of these measurements. (ii) Our Markovian-type model for bipartite graphs ensures transparent propagation of the uncertainties and reproducible results. (iii) The problem of peptides mapping to several protein sequences (often neglected in other models) is addressed automatically according to our statistical model. The performance of our model is shown on three control datasets and compared to the results of two common approaches for protein quantification: APEX [1] and "top3" [2].

Model

Notation



Furthermore we use two "distance" measures

- d_i is the number of proteins having a common edge with peptide i
- d_{ik} is the number of proteins having a common edge with peptide i and peptide k

and define $\underline{U}_r$ as the vector of intensities of all peptides in the connected component r; $\mathfrak{Sigma}_{\underline{U}_r} = \mathrm{Cov}(\underline{U}_r)$ and $\underline{\alpha} = \alpha(1, \ldots, 1)^\intercal$.

Markovian-Type Assumption

Peptides belonging to the same connected component are independent given their matching proteins. This implies that dependencies among peptides are exclusively due to their common proteins. Furthermore, we make a Markovian assumption (for graphical models) which states that only the neighboring proteins matter in the conditional distribution for the peptides (see also [3]).

Model

$C_1, C_2, \ldots, C_{m_r}$ are i.i.d. with $\mathbf{E}[C_j] = 0$ and $\mathrm{Var}(C_j) = 1$.

We propose the following model for the peptide intensities:

$$U_i = \alpha + p_i \beta \, d_i^{-\frac{1}{2}} \sum_{j \in Ne(i)} C_j + \epsilon_i$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_{n_r}$ are i.i.d. with $\mathbf{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \tau^2$. The elements of the covariance matrix of U are then given by

$$\mathrm{Cov}(U_i, U_k) = \begin{cases} p_i \, p_k \, \beta^2 \frac{d_{ik}}{\sqrt{d_i}\sqrt{d_k}} & i \neq k \\ p_i^2 \, \beta^2 + \tau^2 & i = k \end{cases}$$

and the covariance between C_j and U_i is

$$\mathrm{Cov}(C_j, U_i) = \begin{cases} 0 & \text{if there is no edge between } i \text{ and } j \\ p_i \, \beta \frac{1}{\sqrt{d_i}} & \text{if there is an edge between } i \text{ and } j \end{cases}$$

Predicting the protein concentrations

Assume we are working on the first connected component, then the corresponding protein concentrations are given by

$$\mathbf{E}[C_j | \underline{U}_1] = (\underline{U}_1 - \underline{\alpha})^\intercal \, \mathfrak{Sigma}_{\underline{U}_1}^{-1} \begin{pmatrix} \mathrm{Cov}(C_j, U_1) \\ \mathrm{Cov}(C_j, U_2) \\ \vdots \\ \mathrm{Cov}(C_j, U_{n_1}) \end{pmatrix}$$

Parameter estimation

Maximum likelihood estimation (MLE)

$$\underline{U}_r \sim \mathcal{N}_{n_r}(\underline{\alpha}, \mathfrak{Sigma}_{\underline{U}_r})$$

$$f(\underline{U}_r; \alpha, \beta, \tau^2) = |2\pi \mathfrak{Sigma}_{\underline{U}_r}|^{-1/2} \exp\left(-\frac{1}{2}(\underline{U}_r - \underline{\alpha})^\intercal \, \mathfrak{Sigma}_{\underline{U}_r}^{-1} (\underline{U}_r - \underline{\alpha})\right)$$

The negative log-likelihood can then be written as

$$-\sum_{r=1}^{R} \log\left(f(\underline{U}_r; \alpha, \beta, \tau^2)\right)$$

which has to be minimized w.r.t. $\alpha, \beta, \tau^2 > 0$.

Least squares approach (LSA)

Estimate α (mean) and the covariance matrix ($\widehat{\mathfrak{Sigma}}_{\underline{U}_r}$) from the data. Use the off-diagonal elements of $\widehat{\mathfrak{Sigma}}_{\underline{U}_r}$ to estimate β:

$$\sum_{r=1}^{R} \sum_{\substack{i \neq k \\ i,k \in cc_r}} \left(\left(\widehat{\mathfrak{Sigma}}_{\underline{U}_r}\right)_{ik} - p_i \, p_k \, \beta^2 \frac{d_{ik}}{\sqrt{d_i}\sqrt{d_k}} \beta^2 \right)^2 \stackrel{!}{=} \underset{\text{w.r.t. } \beta^2}{\text{minimize}}$$

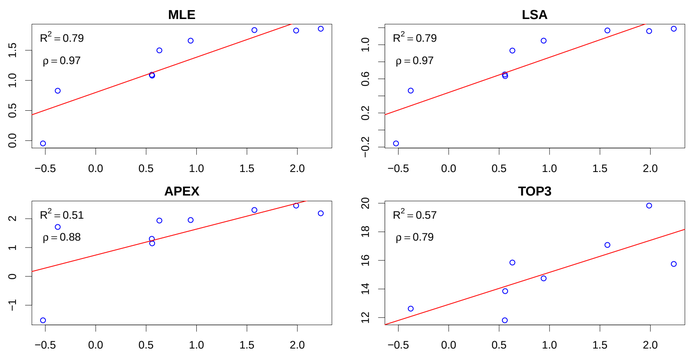
With the diagonal elements and $\hat{\beta}$ one can compute the estimate for τ^2:

$$\sum_{r=1}^{R} \sum_{i=1}^{n_r} \left(\left(\widehat{\mathfrak{Sigma}}_{\underline{U}_r}\right)_{ii} - p_i^2 \, \hat{\beta}^2 - \tau^2 \right)^2 \stackrel{!}{=} \underset{\text{w.r.t. } \tau^2}{\text{minimize}}$$

Results

9 spiked proteins [4]

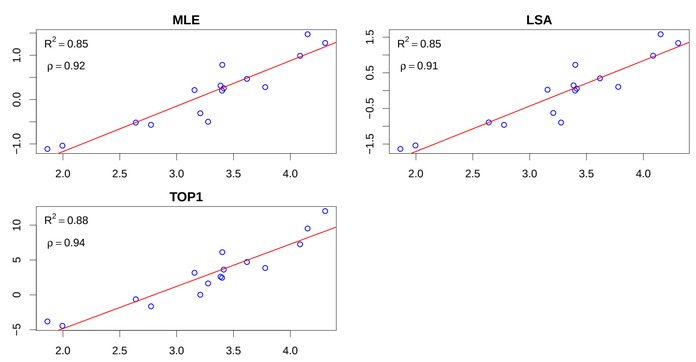
Shotgun (Orbitrap) experiments on 9 non-human proteins spiked into human (K562) cell lysate in 3 different concentrations and analyzed in 6 injections each. We compare the performance of the different methods in one of the mixtures (all technical replicates combined).



computed scores versus known concentrations

Leptospira interrogans [5]

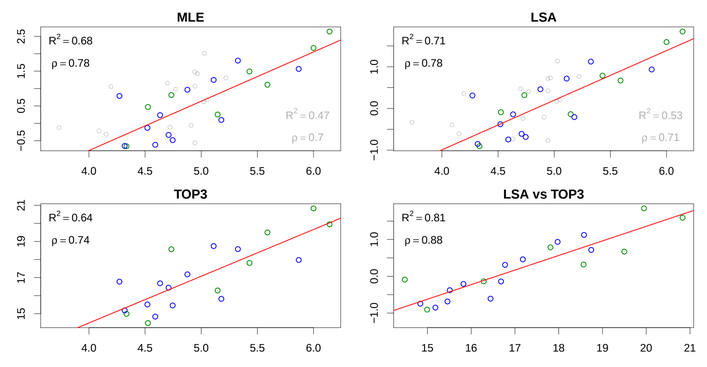
Selected reaction monitoring (SRM) experiment on 16 *L. interrogans* proteins under 3 conditions (with 3 technical replicates each). The proteins were experimentally quantified using AQUA peptides [6]. We compare the performance of the different methods for one of the conditions (all technical replicates combined).



computed scores versus known concentrations

Human shotgun data set [7]

Shotgun experiment (18 OGE fractions) on Human samples under various conditions. 44 proteins were experimentally quantified using AQUA peptides [6]. We compare the performance of different quantification methods on the control sample.
Green symbols correspond to proteins identified (partly) with shared peptides. Grey symbols represent proteins that could not be quantified with the top3 approach.



computed scores versus known concentrations

Outlook & Implementation

Conclusions

The results from our model are competitive with other widely used approaches for protein quantification.
The main advantage of using the least squares approach is to save computation time. It is also possible to use it combined with the MLE in order to get good starting values for the numerical otpimization.
Our model is not designed to work with a particular setting/machinery, but can handle different types of intensity measures for the peptides.

Outlook

The presented approach has a similar performance as the existing tools. Also, it potentially holds two advantages:

- Our model deals *per se* with shared peptides (instead of discarding them) and might thus bring further insight for organisms with an important amount of shared peptides.
- Our method does not rely on spectral counts nor on the proportion of seen versus unseen peptides, and can thus also be used in directed MS experiments or targeted proteomics.

R Code

The presented results were produced R with the following packages/program versions:

- R version 2.13.1 (2011-07-08), `x86_64-unknown-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: sfsmisc 1.0-14, xtable 1.5-6

References

[1] Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotech*, 25(1):117 – 124, 2007.

[2] Jeffrey C. Silva, Marc V. Gorenstein, Guo-Zhong Li, Johannes P. C. Vissers, and Scott J. Geromanos. Absolute quantification of proteins by LCMSE: A virtue of parallel ms acquisition. *Molecular & Cellular Proteomics*, 5:144 – 156, 2006.

[3] Sarah Gerster, Ermir Qeli, Christian H. Ahrens, and Peter Bhlmann. Protein and gene model inference based on statistical modeling in k-par tite graphs. *Proceedings of the National Academy of Sciences*, 107(27):12101 – 12106, 2010.

[4] Christine Vogel. MS/MS shotgun proteomics data repository. `http://www.marcottelab.org/MSdata/`.

[5] Christina Ludwig, Manfred Claassen, Alexander Schmidt, and Ruedi Aebersold. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *publication in progress*, 2011.

[6] Scott A. Gerber, John Rush, Olaf Stemman, Marc W. Kirschner, and Steven P. Gygi. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences*, 100(12):6940 – 6945, 2003.

[7] Martin Beck, Alexander Schmidt, Johan Malmstroem, Manfred Claassen, Alessandro Ori, Anna Szymborska, Franz Herzog, Oliver Rinner, Jan Ellenberg, and Ruedi Aebersold. The quantitative proteome of a human cell line. *submitted*, 2011.

ISMB/ECCB in Vienna, Austria, July 17-19, 2011