Statistical approach to protein quantification

Sarah Gerster[†,§], Christina Ludwig[‡], Mariette Matondo[‡], Ruedi Aebersold[‡] & Peter Bühlmann[†]

ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

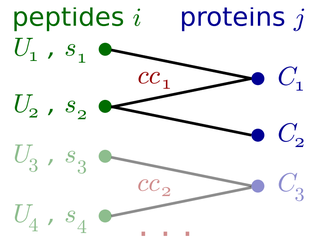
[†]Seminar for Statistics, [‡]Institute of Molecular Systems Biology
[§]current address: SIB Swiss Institute of Bioinformatics – BCF, Lausanne

Abstract

Proteomics provides additional insights into biological systems that cannot be provided by genomic or transcriptomic approaches [1]. In particular, proteomics holds great promise for the identification of biomarkers capable of predicting disease already at a very early stage. For this, accurate identification and quantification of proteins is required. The presented statistical approach to protein quantification, SCAMPI, relies on experimentally identified and quantified peptides. It has six main advantages compared to most existing tools: (i) Peptide abundances are modeled as random quantities, allowing to account for the uncertainty of these measurements. (ii) A 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 is addressed automatically according to our statistical model. (iv) Various types of input data (e.g. shotgun or SRM; labeled or unlabeled) can be handled. (v) The model can be used to reassess the peptide abundance measurements. (vi) A prediction interval is readily provided for each estimated protein abundance score.

Model

Notation



U_i : abundance score *(given)*
s_i : identification score *(given)*
C_j : abundance *(unknown)*
⇒ latent variable
cc_r : connected component with
- n_r peptides
- m_r proteins

$\underline{U}^{(r)}$ is the vector of intensities of all peptides in connected component r. $Ne(i)$ denotes the set of proteins having an edge with peptide i. $\underline{1}^{(r)}$ is a vector of ones of length n_r.

Matrix $D^{(r)}$ ($n_r \times n_r$) holds the connectivity information for cc_r:

→ D_{ii} = number of proteins sharing an edge with peptide i

→ D_{ik} = number of proteins sharing an edge with peptide i and peptide k

Markovian-type assumption

- peptides belonging to the same connected component are independent given their matching proteins

  → dependencies among peptides are exclusively due to their common proteins

- only neighboring proteins matter in the (conditional) distribution for the peptides (see also [2])

Model

We propose the following model for the peptide abundances:

$$U_i \;=\; \alpha \;+\; s_i\,\beta \sum_{j \in Ne(i)} C_j \;+\; \epsilon_i \;, \quad \text{with}$$

$$C_1,\, C_2,\, \ldots,\, C_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu,\, 1\right)$$

$$\epsilon_1,\;\; \epsilon_2,\;\; \ldots,\;\; \epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \tau^2\right)$$

where $\epsilon_1,\;\; \epsilon_2,\;\; \ldots,\;\; \epsilon_n$ are independent of $C_1,\, C_2,\, \ldots,\, C_m$.

The elements of the covariance matrix of $\underline{U}$ are then given by

$$\mathrm{Cov}\left(U_i, U_k\right) \;=\; \left(\mathbb{\Sigma}_{\underline{U}^{(r)}}\right)_{ik} \;=\; \begin{cases} s_i^{(r)}\; s_k^{(r)}\; \beta^2\; D_{ik}^{(r)} & \text{for } i \neq k \\ \left(s_i^{(r)}\right)^2\; \beta^2\; D_{ii}^{(r)} \;+\; \tau^2 & \text{for } i = k \end{cases}$$

and the covariance between C_j and U_i is

$$\mathrm{Cov}\left(C_j, U_i\right) = \left(\Gamma_{C_j \underline{U}^{(r)}}\right)_i = \begin{cases} 0 & \text{for } j \notin Ne(i) \\ s_i^{(r)}\; \beta & \text{for } j \in Ne(i) \end{cases}$$

Predicting protein abundances

For protein j in connected component r:

$$\widehat{C}_j = \mathbf{E}\left[C_j | \underline{U}^{(r)}\right] = \mu + \left(\underline{U}^{(r)} - \alpha \underline{1}^{(r)} - \underline{s}^{(r)}\,\beta\,\mu\;\mathrm{diag}(D^{(r)})\right)^{\intercal}\; \mathbb{\Sigma}_{\underline{U}^{(r)}}^{-1}\; \Gamma_{C_j \underline{U}^{(r)}}$$

$$\mathrm{Var}\left(C_j | \underline{U}^{(r)}\right) = 1 - \Gamma_{C_j \underline{U}^{(r)}}^{\intercal}\; \mathbb{\Sigma}_{\underline{U}^{(r)}}^{-1}\; \Gamma_{C_j \underline{U}^{(r)}}$$

Peptide reassessment

Protein abundance scores can be used to estimate peptide abundances. These values $\widehat{U}_i$ can then be compared to the measurements U_i to detect discrepancies in the input data. To avoid overfitting, the expected value for the abundance of peptide i given all other peptide measurements ($\{U_{k \setminus i}\}$) is computed:

$$\widehat{U}_i \;=\; \mathbf{E}\left[U_i | \{U_{k \setminus i}\}\right] \;=\; \alpha \;+\; s_i\,\beta \sum_{j \in Ne(i)} \mathbf{E}\left[C_j | \{U_{k \setminus i}\}\right]$$

Parameter estimation

Method of moments approach relying on least squares estimates on the elements of the covariance matrix $\mathbb{\Sigma}$:

1. estimate α and $\beta\mu$ by fitting a linear regression $U \sim s\,\mathrm{diag}(D)$
2. use sample covariance matrix of U to estimate β and τ

   → off-diagonal elements of $\widehat{\mathbb{\Sigma}}_{\underline{U}^{(r)}}$ allow to estimate β:

$$\sum_{r=1}^{R} \sum_{\substack{i \neq k \\ i,k \in cc_r}} \left(\left(\widehat{\mathbb{\Sigma}}_{\underline{U}^{(r)}}\right)_{ik} - s_i^{(r)}\, s_k^{(r)}\, D_{ik}^{(r)}\, \beta^2\right)^2 \overset{!}{=} \underset{\text{w.r.t. } \beta^2}{\text{minimize}}$$

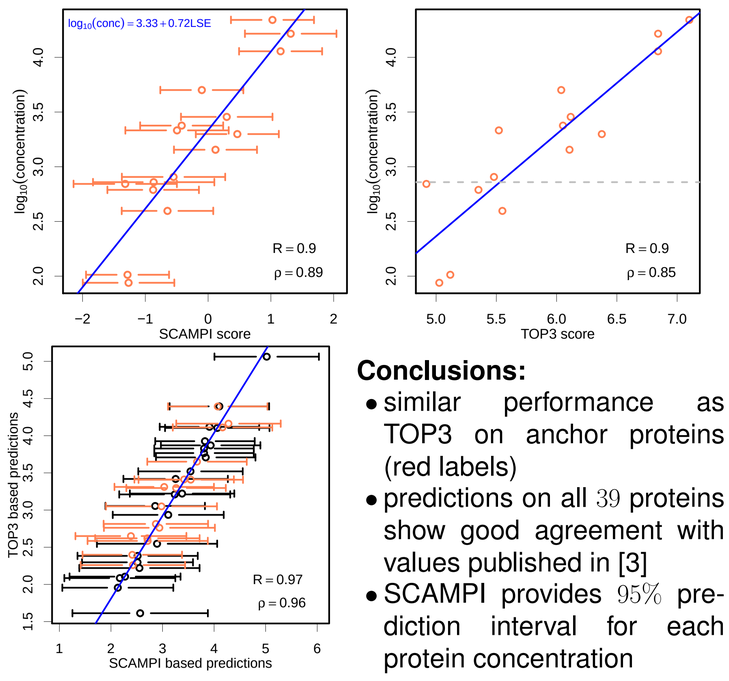
   → diagonal elements of $\widehat{\mathbb{\Sigma}}_{\underline{U}^{(r)}}$ and $\hat{\beta}$ yield an estimate for τ^2:

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

Results

Leptospira interrogans [3]

Selected reaction monitoring (SRM) experiment on 39 *L. interrogans* proteins under 3 conditions (with 3 technical replicates each). 16 (anchor) proteins were experimentally quantified using AQUA peptides [4]. We compare the performance of SCAMPI and TOP3 [5] for the control condition (all technical replicates combined). Performance is measured in terms of Pearson's correlation coefficient (R) and Spearman's rank correlation coefficient (ρ).



Conclusions:

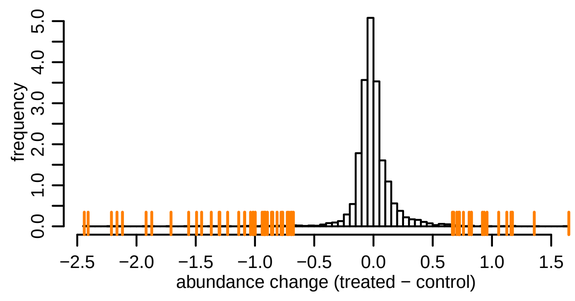
- similar performance as TOP3 on anchor proteins (red labels)
- predictions on all 39 proteins show good agreement with values published in [3]
- SCAMPI provides 95% prediction interval for each protein concentration

SILAC-labeled human shotgun data [6]

SILAC labeled data from a human acute myeloid leukaemia cell line (KG1a cells). We compare the protein expressions in the control to the treatment condition for the cytoplasmic fraction. Ground truth is not known, but the data allows to show how SCAMPI can do relative quantification and reassess peptide measurements in datasets with a large percentage of shared peptides (about 20%).

Relative protein quantification

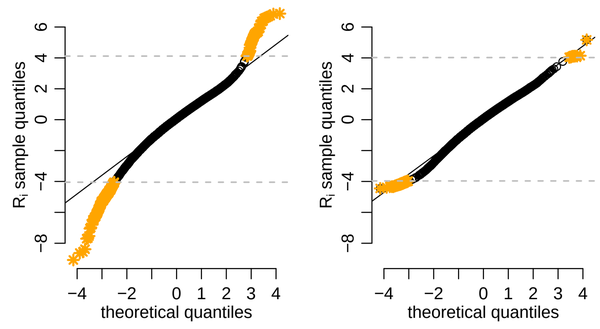
Proteins are quantified separately for control and treatment, respectively. Score differences are considered to identify the proteins undergoing the most important changes in abundance between the two conditions. Proteins with a particularly high score difference, namely with a corrected p-value smaller than 5%, are shown in orange in the plot below.

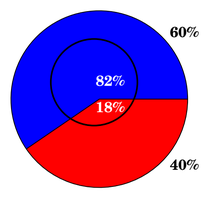


Proof of principle: some of the proteins found to be significantly differentially abundant belong to the HSP family, and have been reported in other publications as well.

Peptide intensity outlier detection

Peptide reassessment can be used to recursively improve the protein abundance predictions. Here SCAMPI ran twice. The normal Q-Q plot for the peptide residuals after the first (left) and second run are shown. For the second run, all peptides highlighted in orange in the left plot (about 300 out of 30′000) were removed.



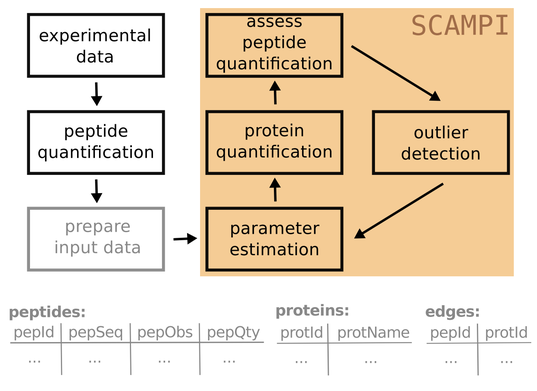


Proof of principle:

- many peptides with underestimated abundance scores contain at least one missed cleavage site
- SCAMPI explains high abundant shared peptides overproportionally well (diagram on the left)

Outlook & Implementation

Typical SCAMPI workflow



Reasons to use SCAMPI

- information in shared peptides is used
- parameters are trained on the whole dataset, even if predictions are only required for a subset of the proteins
- prediction interval is provided for the computed protein abundance scores
- relative quantification relies on a statistical test with correction for multiple testing and stringent cutoff
- R code is available upon request, and soon also as an R package on CRAN

R Code

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

- R version 2.15.1 (2012-06-22), `x86_64-unknown-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils
- Other packages: KernSmooth 2.23-8, MASS 7.3-20, RBGL 1.32.1, bitops 1.0-4.1, caTools 1.13, gdata 2.11.0, gplots 2.11.0, graph 1.34.0, gtools 2.7.0, mvtnorm 0.9-9992
- Loaded via a namespace (and not attached): BiocGenerics 0.2.0, tools 2.15.1

References

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[2] Sarah Gerster, Ermir Qeli, Christian H. Ahrens, and Peter Bühlmann. 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.

[3] Christina Ludwig, Manfred Claassen, Alexander Schmidt, and Ruedi Aebersold. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Molecular & Cellular Proteomics*, 11(3), 2012.

[4] 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.

[5] 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.

[6] Mariette Matondo. Unpublished SILAC experiments on KG1a cells.