STATISTICAL APPROACH TO PROTEIN QUANTIFICATION

ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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

†Seminar for Statistics, ‡Institute of Molecular Systems Biology
gerster@stat.math.ethz.ch

Model

Abstract

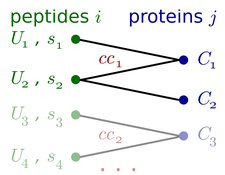
A major goal in proteomics is the comprehensive and accurate description of a proteome. Proteomics provides additional insights into biological systems that cannot be provided by genomic or transcriptomic approaches [1]. In particular, proteomics holds great promise for the identification and quantification of biomarkers capable of accurately predicting diseases already at a very early stage.

We propose a statistical approach to protein quantification based on shotgun experiments with four main advantages:

(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.
(iv) Peptides with outlying intensity values can be assessed and classified as either regular data points or "true" outliers.

The application of our model is illustrated on two data sets and the performance is compared to another common approach for protein quantification [2].

Notation



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

We define $\underline{U}^{(r)}$ as the vector of intensities of all peptides in the connected component r. By writing $\underline{\alpha}$ we mean $\alpha\,(1,\ldots,1)^\intercal$.

Furthermore, we use a "distance" matrix D with
- D_{ii} = number of proteins having a common edge with peptide i
- D_{ik} = number of proteins having a common edge with peptides i and k

Markovian-type assumptions

- 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 [3])

Model

We propose the following model for the peptide intensities:

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

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

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

The elements of the covariance matrix of U are then given by

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

and the covariance between C_j and U_i is

$$\text{Cov}\left(C_j, U_i\right) = \Gamma_{C_j U_i} = \begin{cases} 0 & \text{if there is } \textit{no} \text{ edge between } i \text{ and } j \\ s_i\;\beta & \text{if there is an edge between } i \text{ and } j \end{cases}$$

Predicting protein concentrations

For a connected component r:

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

Parameter estimation

Maximum likelihood estimation (MLE)

$$\underline{U}^{(r)} \sim \mathcal{N}_{n_r}\left(\underline{m},\; \mathfrak{Sigma}_{\underline{U}^{(r)}}\right) \quad \text{with} \quad \underline{m} = \underline{\alpha} + \underline{s}^{(r)}\,\beta\,\mu\;\text{diag}(D^{(r)})$$

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

Minimize negative log-likelihood with respect to α, β, μ and τ^2

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

Least squares approach (LSA)

Estimate α and $\beta\mu$ by fitting $U \sim s \cdot \text{diag}(D)$. Use sample covariance matrix of U to estimate β and τ.
Off-diagonal elements of $\widehat{\mathfrak{Sigma}}_{\underline{U}^{(r)}}$ allow to estimate β:

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

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

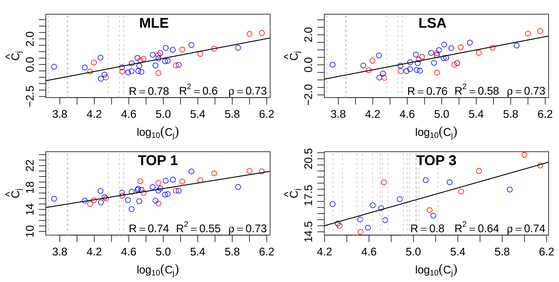
$$\sum_{r=1}^{R} \sum_{i=1}^{n_r} \left(\left(\widehat{\mathfrak{Sigma}}_{\underline{U}^{(r)}}\right)_{ii} - s_i^2\, \hat{\beta}^2\, D_{ii} - \tau^2\right)^2 \overset{!}{=} \underset{\text{w.r.t. } \tau^2}{\text{minimize}}$$

Results

Human shotgun data [4]

"Absolute" protein quantification

Shotgun experiment (18 OGE fractions) on Human samples under various conditions. Performances of different quantification methods are compared on 44 experimentally quantified proteins (using AQUA peptides [5]) in the control sample.
Red symbols correspond to proteins identified (partly) with shared peptides. Grey lines indicate ground truth proteins that could not be quantified.



computed scores versus known concentrations

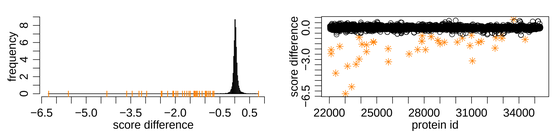
SILAC-labeled human shotgun data [6]

Comparison between a control sample and irradiated (SILAC-labeled) human cells. The data sets contain much more shared peptides than unique ones. The bipartite graph holds:

- 22040 peptide and 13315 protein sequences
- 16728 of the peptides match to several proteins (shared)
- 12743 proteins match to at least one shared peptide
- 12018 proteins have no unique peptide evidence
- only 1297 proteins can be quantified with `TOP1` (560 with `TOP3`)

Relative protein quantification

Proteins are quantified separately for the control and treatment sample, respectively. Score differences are used to find the proteins with the most important concentration changes between the two samples. Proteins with particularly high score differences are shown in orange.

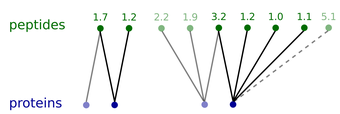


protein concentration score differences: "treatment - control"

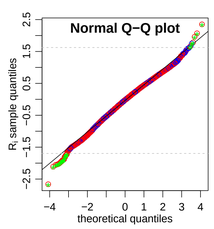
Peptide intensity outlier detection

There are various reasons to encounter outliers in the measured peptide intensities:

- measurement errors
- shared peptides
- modifications
- missed cleavages
- incomplete database



Predicting peptide intensities and analyzing the residuals can potentially lead to further insight about the data. Monitoring peptides with large intensity measurements which could *not* be explained by the model (high residuals) will for example help to assess how useful the new approach is when dealing with shared peptides.



	shared	unique	ratio
all data	16728	5312	3.15
20%	1511	519	2.91
10%	653	241	2.71
5%	284	111	2.56
1%	36	16	2.25

Interpretation: the line labeled with 5% reports the number of shared and unique peptides belonging to the set of the top 5% of the measured intensities *and* to the 2.5% lowest or to the 2.5% highest residuals.

Conclusion & outlook

We provide a simple model with clearly stated assumptions for protein quantification.

- performance similar to other approaches
- when working on data sets with shared peptides our model allows to quantify all proteins
- potential to get further insight on the peptide level (assessing intensity measurements)
- flexible (peptide scores, input type, experiment type,...)

Implementation

Our model is implemented in R [7]. The presented results are obtained with the following version:

- R version 2.15.0 RC (2012-03-25 r58832), `x86_64-unknown-linux-gnu`
- Base packages: base, data sets, graphics, grDevices, methods, splines, stats, utils
- Other packages: MASS 7.3-17, regr0 1.0-2, sfsmisc 1.0-20, xtable 1.7-0
- Loaded via a namespace (and not attached): tools 2.15.0

References

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[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] 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. *Molecular Systems Biology*, 7(549), 2011.

[5] 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.

[6] Mariette Matondo. Silac-labeled human shotgun data. personal communication.

[7] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.