Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-26 (2016) 056–062

Normalization of Western blot data affects the statistics of estimators [★]

Caterina Thomaseth [*] **Nicole Radde** [*]

[*] *Institute for Systems Theory and Automatic Control, University of Stuttgart, 70569 Stuttgart, Germany (e-mails: caterina.thomaseth@ist.uni-stuttgart.de, nicole.radde@ist.uni-stuttgart.de).*

Abstract: This study elaborates on the normalization of data from Western blot experiments and its impact on parameter estimation. Western blot data have to be preprocessed appropriately in order to enable comparison across different replicates. This includes a two step normalization procedure, in which the raw signals are normalized to a loading control and additionally to a reference condition. If the signals themselves are normally distributed, the normalized data are described by ratios of normal distributions, which have some peculiarities that can complicate further analysis such as parameter estimation for biochemical network reconstruction. Here we shortly recapitulate some properties of these ratio distributions and conditions for various approximations that facilitate further analysis. We illustrate results on a case study in which Western blot data are used to infer the fold change in a knockdown experiment.

Keywords: Estimators, Likelihood function, Maximum likelihood principle, Monte Carlo simulation, Ratio distribution

1. INTRODUCTION

Western blotting is a technique for the quantification of proteins' concentrations or their phosphorylation states, which is in use since about three decades. Proteins isolated from a cell culture are transferred to a membrane and incubated with a specific primary antibody. Then the membrane is incubated with a secondary antibody, which is directed against the first antibody and serves as signal amplifier. This second antibody is chemically linked to an enzyme that catalyzes a chemiluminescence reaction. Finally, exposure of an X-ray film produces bands, indicating the location of a protein-antibody complex. In the linear range, the band intensity is proportional to the amount of the protein. Since proportionality factors are membrane and antibody specific, normalization has to be performed to enable comparison between different replicates. This is commonly done in a two step procedure, in which signals are first normalized to a loading control to reduce the variance resulting from loading differences among the lanes. In a second step the data are additionally normalized to a control or reference condition. For example, if monitoring the temporal change of the phosphorylation state of a protein in an intracellular signaling cascade upon stimulation with a ligand, the state in the unstimulated case might serve as such a reference condition. Thus, normalized data are given as multiples of this reference state. Similarly, the altered concentration of a protein in a perturbation experiment, e.g. overexpression or knockdown, is normalized to the signal in the unperturbed case. Originally Western blotting was mainly used to detect differences in the amount or activity state of proteins across different conditions, but there is a continuous trend to extract also more quantitative information. Due to improvements in experimental protocols to determine linear ranges between optical densities and protein concentrations and to perform proper background corrections, and further developments regarding the experimental technique (see e.g. Taylor and Posch (2014)), Western blot data are nowadays also more and more frequently used for parameter estimation of quantitative models (see e.g. Weber et al. (2015)).

Accordingly, research has started towards characterizing the statistical properties of Western blot data (Degasperi et al., 2014; Kreutz et al., 2007). Although often not stated explicitly, already the representation of different replicates as summary statistics such as a mean and a variance and their calculation make an implicit assumption about underlying distributions and statistical properties of the data. A common assumption for representation purposes and also for hypothesis testing are normally distributed data or summary statistics. Examples are the t-test or ANOVA to determine the significance of differences in mean values (see e.g. Möller et al. (2014); Zinöcker and Vaage (2012)). On the other hand, in Kreutz et al. (2007) it is argued, based on a comparison of different error models, that Western blot signals rather follow a log normal distribution.

In this study we consider these two commonly used statistical models, normal and log normal distributions, for

[★] This work was supported by the German Federal Ministry of Education and Research (BMBF) within the e:Bio-Innovationswettbewerb Systembiologie project PREDICT (grant number FKZ0316186A) and the German Research Foundation (DFG) within the Cluster of Excellence in Simulation Technology (EXC 310/2) at the University of Stuttgart.

10.1016/j.ifacol.2016.12.103

Western blot data. These distribution assumptions have very different implications on the distributions of the normalized data, which will be discussed. Moreover, the assumptions on the underlying distribution, or on the error model, have an impact on parameter identification, which is demonstrated here for maximum likelihood estimators. We exemplify results on a real-world case study, in which a protein is quantified in the control case and in a knockdown experiment (Santos et al., 2007), and data are used to determine the fold change.

2. STOCHASTIC MODELING FRAMEWORK

A stochastic modeling framework describes data y as random variates that are generated by an underlying stochastic process. This process defines the distribution $p(y)$. Statistical learning deals with the problem of inference of $p(y)$ from observations. Here we consider the case of distributions that are parametrized by parameters θ, such that statistical learning is equivalent to the estimation of these parameters from observations y. We consider three different stochastic scenarios for the description of a knockdown experiment.

2.1 Statistical description of a knockdown experiment

We consider a Western blot dataset from a knockdown experiment, in which the amount of protein has been quantified under control and knockdown conditions. In order to estimate the fold change, i.e. the factor by which the protein amount is reduced compared to the amount in the control case, the data of the knockdown experiments are commonly presented normalized to the control case. Here we consider three different scenarios of the stochastic process underlying data generation:

Scenario 1: No variance in the control experimental data
This is the simplest scenario, in which data x_c of the control experiments are assumed to be errorless, while data x_k from the knockdown experiments are assumed normally distributed:

$$\begin{aligned} X_c &= \delta(\mu_c) \\ X_k &\sim \mathcal{N}(\mu_k, \sigma_k^2) \end{aligned} \tag{1}$$

Scenario 2: Control and knockdown experimental data are log normally distributed
If X_c and X_k follow a log normal distribution, their logarithms are normally distributed:

$$\begin{aligned} logX_c &\sim \mathcal{N}(\mu_c, \sigma_c^2) \\ logX_k &\sim \mathcal{N}(\mu_k, \sigma_k^2) \end{aligned} \tag{2}$$

Scenario 3: Control and the knockdown experimental data are normally distributed
Here we consider the case

$$\begin{aligned} X_c &\sim \mathcal{N}(\mu_c, \sigma_c^2) \\ X_k &\sim \mathcal{N}(\mu_k, \sigma_k^2) \end{aligned} \tag{3}$$

We will consider two subcases of scenarios 2 and 3, in which X_c and X_k are independently distributed, i.e. Cov(X_c, X_k) $= \rho = 0$ (scenarios 2A and 3A), and the probably more realistic case in which X_c and X_k are correlated, $\rho \neq 0$ (scenarios 2B and 3B).

The variables X_c and X_k describe protein amounts, which cannot be observed directly in Western blots. Instead, optical densities from secondary antibodies are measured, which are assumed to be proportional to these amounts, with membrane and antibody specific constants α_j for each replicate. Hence the distributions of signals $\tilde{Y}_c^j$ and $\tilde{Y}_k^j$ that describe the observed optical densities in each replicate $j = 1, \dots J$ are given by

$$\begin{aligned} \tilde{Y}_c^j &= \alpha_j X_c \\ \tilde{Y}_k^j &= \alpha_j X_k. \end{aligned} \tag{4}$$

Finally, in order to get rid of the prefactors α_j, all replicates are normalized to the control condition, which gives normalized data y as

$$Y_c = \frac{\tilde{Y}_c^j}{\tilde{Y}_c^j} = 1 \, \forall j \quad \text{and} \quad Y_k = \frac{\tilde{Y}_k^j}{\tilde{Y}_c^j}. \tag{5}$$

Table 1 lists the respective distributions of variables $\tilde{Y}$ and Y for all three scenarios. For scenarios 1 and 2 the distributions of Y_k belong to the same class of distributions as assumed for the original concentrations X_k, namely normal and log normal distributions. Mean and variance are sufficient statistics for these distributions. In scenario 3, Y_k is described by a ratio distribution of two normal random variables. This family of distributions does not belong to the family of exponential distributions (for details on this class see Gelman et al. (2004)), and can show some pecularities that are recapitulated in the following.

2.2 Statistical properties of ratio distributions of normal random variables

Ratio distributions $Z = X/Y$ of two normal random variables $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and correlation ρ have intensively been studies in the 60ies and 70ies, see e.g. Marsaglia (1965); Hinkley (1969); Hayya et al. (1975), with some refined work described in Marsaglia (2006). It was shown in Marsaglia (1965, 2006) that after rescaling and translation, the distribution of $\tilde{Z} = r(Z - s)$ equals those of $T = \frac{a+V}{b+W}$, with W, V independently standard normally distributed and $r, s, a, b \geq 0$ defined as

$$\begin{aligned} r &= \frac{\sigma_Y}{\pm\sigma_X\sqrt{1-\rho^2}} & s &= \frac{\rho\sigma_X}{\sigma_Y} \\ a &= \pm\frac{\mu_X/\sigma_X - \rho\mu_Y/\sigma_Y}{\sqrt{1-\rho^2}} & b &= \frac{\mu_Y}{\sigma_Y}. \end{aligned} \tag{6}$$

The sign in the equation for a has to be chosen such that a and b have the same sign. The density of T can be written as

$$f_T(t) = \frac{e^{-\frac{1}{2}(a^2+b^2)}}{\pi(1+t^2)}\left[1 + qe^{\frac{1}{2}q^2}\int_0^q e^{-\frac{1}{2}x^2}dx\right], \tag{7}$$

with

$$q = \frac{b + at}{\sqrt{1+t^2}}. \tag{8}$$

The integral makes it a little bit difficult to infer properties of this distribution directly via simple analysis, yet we can