BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 20 2007, pages 2747–2753
doi:10.1093/bioinformatics/btm397

Systems biology

An error model for protein quantification

C. Kreutz[1,2,*,†], M.M. Bartolome Rodriguez[3,†], T. Maiwald[1,2], M. Seidl[3], H.E. Blum[3], L. Mohr[3] and J. Timmer[1,2]

[1]Freiburg Center for Data Analysis and Modeling FDM, Eckerstrasse 1, [2]Department for Physics, University of Freiburg, Hermann-Herder-Strasse 3, D-79104 Freiburg and [3]Department of Medicine II, University Hospital of Freiburg, Hugstetter Strasse 55, D-79106 Freiburg, Germany

Received on May 7, 2007; revised on July 11, 2007; accepted on August 2, 2007

Advance Access publication September 3, 2007

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Quantitative experimental data is the critical bottleneck in the modeling of dynamic cellular processes in systems biology. Here, we present statistical approaches improving reproducibility of protein quantification by immunoprecipitation and immunoblotting.

Results: Based on a large data set with more than 3600 data points, we unravel that the main sources of biological variability and experimental noise are multiplicative and log-normally distributed. Therefore, we suggest a log-transformation of the data to obtain additive normally distributed noise. After this transformation, common statistical procedures can be applied to analyze the data.

An error model is introduced to account for technical as well as biological variability. Elimination of these systematic errors decrease variability of measurements and allow for a more precise estimation of underlying dynamics of protein concentrations in cellular signaling.

The proposed error model is relevant for simulation studies, parameter estimation and model selection, basic tools of systems biology.

Availability: Matlab and R code is available from the authors on request. The data can be downloaded from our website www.fdm.uni-freiburg.de/~ckreutz/data.

Contact: ckreutz@fdm.uni-freiburg.de

1 INTRODUCTION

Studies of protein abundance by immunoblotting have been used widely for biological as well as biochemical investigations (Kurien and Scofield 2006). Immunoblotting allows analysis of protein concentrations in cell populations without enhancing their basal expression even for low abundant proteins. Additionally, post-translational modifications, e.g. phosphorylation, can be quantified by immunoblotting. These modifications are crucial for biological functions of proteins in signaling pathways. Quantitative analysis of protein phosphorylation dynamics is therefore essential for the development of systems biological approaches.

Unfortunately, immunoblotting displays a minor signal-to-noise ratio and it is difficult to obtain reproducible quantitative measurements. A further source of noise is biological variability, especially if, e.g. primary cells are used for experiments. Another problem is that the common assumption of normally distributed noise is strongly violated.

Error models describe the distribution of measurements given experimental parameters. Error models can be used to detect and correct systematic sources of noise in experimental approaches. Such error reduction procedures are very common for some other experimental technologies. For example, it was shown that microarray data should be transformed to fulfill desired statistical properties (Durbin and Rocke 2003; Huber *et al*., 2002; Rocke and Durbin, 2003). It has been shown that log-transformed microarray intensities have to be corrected for systematic errors (Bolstad *et al*., 2003; Yang *et al*., 2002) and that error models can be used to estimate differential expression (Ideker *et al*., 2000; Pavelka *et al*., 2004). Statistical analysis of microarray intensities has been studied extensively. A next challenging but necessary step is the development of accurate statistical data processing for other experimental techniques.

Determination of an appropriate error model covers the following decisions:

- Which background correction procedure is appropriate?
- Which transformation should be applied to background corrected measurements?
- Which sources of systematic errors exist and should therefore be accounted in an error model?
- Which systematic errors can be modeled as normally distributed random variable?

Since these issues depend on each other, they cannot be answered separately. Therefore, combinations of transformation, background correction and systematic errors had to be considered.

We investigated a set of 26 error models. The finally superior model for log-transformed immunoblotting intensities accounts for biological as well as experimental noise and shows normally distributed residuals. It will be shown that this model improves reproducibility and increases signal-to-noise ratio by more than a factor of 10 compared to raw background subtracted intensities.

Additionally, it is demonstrated how error models are extended to estimate time dependency of protein concentrations

*To whom correspondence should be addressed.
†The authors wish it to be known that, in this opinion, the first two authors should be regarded as joint First Authors.

and their confidence intervals after stimulation. Despite a large amount of measurement noise, this approach leads to reliable time courses which can now be further evaluated by dynamic modeling approaches.

2 EXPERIMENTAL DATA

For our study, we use measurements of 12 observables, e.g. housekeeping proteins and proteins involved in the insulin signaling pathway. Time courses of activated (indicated by *) insulin receptor (IR*) and insulin receptor substrate (IRS-1*) as well as binding of phosphoinositide kinase (PI3K) to IRS-1 and activation of extracellular regulated kinase (ERK-1* and ERK-2*) as functional outcome of the insulin are measured for different insulin stimulations.

Additionally, total IR, IRS-1, ERK-1 and ERK-2 concentrations as well as some housekeeping proteins (glycoprotein gp96, cellular heat shock cognate hsc70, β-actin) are measured. Further, one positive control of total activation after pervanadate stimulation and one negative control without addition of insulin are performed within each stimulation with insulin. Total concentrations and housekeeping proteins are not affected by insulin stimulation and are therefore considered to be constant. Altogether 3642 data points are analyzed comprising 2108 measurements of time depending events, 1123 measurements of housekeeping proteins and 411 controls.

A cell preparation consists of primary hepatocytes obtained from two mice livers. On each gel, 10–20 different probes are measured in adjacent lanes. Probes are loaded randomized on gels, e.g. not in chronological order (Schilling *et al.*, 2005a). Foreground intensities, as well as background intensities are determined for each spot (see Fig. 1). Details of cell preparation and stimulation are described in the Supplementary Material.

3 METHODOLOGY

In this chapter, a relationship between measurements and underlying protein concentrations is introduced for additive and multiplicative noise. It is discussed how error models are extended to account for systematic measurement errors. Afterwards, methods for comparison and assessment of different error models are introduced.

3.1 Additive and multiplicative errors

Protein concentration as variable of interest cannot be observed directly. Therefore, other quantities y, e.g. fluorescence intensities which are related to variables of interest x are determined experimentally. To avoid saturation effects, experiments are usually optimized to achieve a linear dependency of y on x. Nevertheless, measurements y are always affected by measurement noise. Most common measurement errors are additive. If many independent additive sources ε_i contribute to observed noise $\varepsilon = \sum_i \varepsilon_i$, measurements

$$y = \alpha + \beta x + \varepsilon,\ \varepsilon \sim N(0, \sigma_\varepsilon) \tag{1}$$

are normally distributed with SD σ_ε. Constant offset α represents a systematic shift and β denotes a scaling factor. The variable y on the left-hand side is often called *response variable* of a model. Variables on the right-hand side are called *predictor variables*.

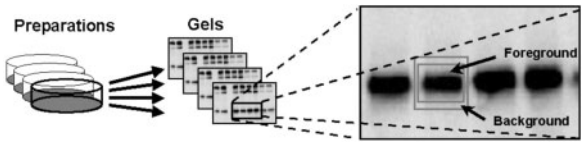

Fig. 1. Each preparation is repeatedly measured on 8–16 gels. Foreground intensities of spots correspond to protein concentrations. Background was determined locally around the spots.

In accordance to Equation (1) most statistical procedures assume a linear relationship between measurements and underlying variable of interest affected by additive Gaussian, e.g. normally distributed errors. If these assumptions are violated, statistical analysis can be refined or, usually the easier way (Atkinson 1981), measurements have to be transformed.

Under weak assumptions, multiplicative noise $\eta = \prod_i \varepsilon_i$ leads to log-normally distributed measurement errors $\eta \sim e^{N(0,\sigma_\eta)}$. Multiplicative noise is often observed for non-negative data, e.g. fluoresence intensities (Limpert *et al.*, 2001).

For measurements $\tilde{y}$ with multiplicative errors, it holds

$$\tilde{y} = \beta_0 + \beta_1 x^{\beta_2} \eta, \eta \sim e^{N(0,\sigma_\eta)}. \tag{2}$$

According to this model, $\tilde{y} - \beta_0$ is log-normally distributed with parameters β_1, β_2 and σ_η. A log-transformation $y = \log(\tilde{y} - \beta_0)$ of this data leads to Equation (1) for $\log(x)$ with $\alpha = \log(\beta_1)$, $\beta = \beta_2$, $\varepsilon = \log(\eta)$ and $\sigma_\varepsilon = \sigma_\eta$.

After such a log-transformation, all statistical methods assuming normally distributed noise can be applied, e.g. averaging can be performed for calculation of expectation values. Additionally, error bars which are asymmetric for observations $\tilde{y}$ become symmetric on log scale.

For immunoblotting, an error model

$$\tilde{y} = \beta_0 + \beta_1 x(1 + \eta), \eta \sim N(0, \sigma_\eta) \tag{3}$$

assuming additive errors with an SD proportional to signal intensities was already introduced (Schilling *et al.*, 2005b; Swameye *et al.*, 2003). For small multiplicative noise $\sigma_\eta \ll x$ this model represents a first-order approximation of the multiplicative error model (2) for special case $\beta_2 = 1$. This approximation shows intensity-dependent error bars. But, in contrast to model (2), error bars are symmetric on measurement scale for model (3).

Recapitulating, in the case of multiplicative noise log-transformation can be used to allow application of statistical methods which assume additive Gaussian errors. The case of commonly occurring additive and multiplicative noise, is discussed in Supplementary Material. Our analyses show that immunoblotting data is dominated by multiplicative noise and that log-transformation is sufficient to obtain Gaussian errors.

3.2 Mixed effects models

Equation (1) describes a linear relationship of possibly transformed observations with true underlying constant protein concentrations x. The model has to be extended for time-dependent proteins or if systematic errors should be accounted. Systematic errors are correlated and usually related to experimental parameters.

Overall systematic errors are considered in Equation (1) by offset α. Additional errors like a preparation or gel-specific

systematic shift would lead to a biased estimate of underlying concentrations x. If experimental sources of such influences are known, they should be accounted by enlarging model (1). Auxiliary new predictor variables, so called '*effects*', can be added at the right hand side of (1) or in the case of multiplicative influences, β_0 in (2) can be replaced by a product of different effects. An 'effect' represents the amount of influence on measurements of one experimental parameter. Fitting a suitable enlarged model to measurements enables an unbiased estimate of underlying concentrations.

Discrete experimental parameters are modeled as *fixed effects*. For example, gel differences can be accounted by introduction of a gel effect G_g. Index $g = 1, \ldots, n_{\text{gels}}$ enumerates different gels and n_{gels} parameters G_g have to be estimated. More than one index indicates that the magnitude of an effect depends on more than one experimental parameter.

For discrete predictor variables, the number of parameters can be reduced if an effect is interpreted as a random variable. In contrast to a fixed effect, a normally distributed *random effect* requires only one parameter for its variance. In our example, random gel effects would be described by a total gel-to-gel variability σ_{gel}. Discrete effects should be modeled by random variables to avoid over-parametrization if the assumption of normal distribution is fulfilled. Statistical models with fixed and random effects are called *mixed effects models* (Pinheiro and Bates 2000). Mixed effects models constitute an established statistical framework for modeling of multiple sources of variation.

Influences of a continuous variable like background intensity are modeled via continuous predictor variables with one regression parameter. In signal transduction, protein concentrations usually depend on time and stimulation treatment. Since the exact functional relationship is unknown, time and stimulation cannot be modeled as continuous predictor variables. For time-dependent protein concentrations, x in Equation (1) has to be replaced by discrete time effects T_{ost}. Index o enumerates different observables, e.g. proteins, $s = 1, \ldots, n_s(o)$ enumerates different stimulation treatment effects and $t = 1, \ldots, n_t(o, s)$ enumerates different times after stimulation. The number of applied stimulations $n_s(o)$ depends on the observable and the number of measured times $n_t(o, s)$ depends on the observable and the applied stimulation. Matrix notation of considered mixed effects models is described in Supplementary Material.

Further, a cell preparation effect P is introduced to account for biological variability and overall difference of observables, e.g. obtained by unequal specificity of antibodies, is modeled via an observable effect O_o. Table 1 displays considered effects for our models. All regarded effects are combined yielding to different error models which are compared with measurements. In the next section, several methods are introduced to decide which effects are required and which effects have to be considered as fixed or random variables. All models are fitted using R statistical software environment (www.r-project.org).

3.3 Assessing required effects

Akaike Information Criterion (Sakamoto *et al.*, 1986) and *Bayes' Information Criterion* (Schwarz *et al.*, 1978) are used to assess the relative fit of competing error models. Akaike

Table 1. Properties of regarded effects

Effect name	Type of effect	Representation in the models
Observable O	Discrete, fixed	Always considered
Preparation P	Discrete, fixed/random	Considered or discarded
Gel number G	Discrete, fixed/random	Always nested within P
Background B	Continuous, fixed	Predictor or response variable
Time effect T	Discrete, fixed	Not required for housekeeping proteins

The error models differ in the way of accounting for background, protein-, preparation- and gel-specific effects.

Information Criterion is defined as

$$\text{AIC} = -2\log(L(y|M)) + 2\,n_{\text{par}} \tag{4}$$

where L is the likelihood function, e.g. the probability of the data y given an error model M with n_{par} parameters. Bayes' Information Criterion

$$\text{BIC} = -2\log(L(y|M)) + n_{\text{par}}\log(n) \tag{5}$$

is similarly defined but takes into account the number of data points n. Small values of these criteria are preferable and are obtained by a large likelihood function and a small number of parameters. Usually, BIC tend to select smaller models than AIC, especially for large number of data points.

Analysis of Variance (ANOVA) is used to check which fixed effects are not capable to explain variance significantly. Such effects should be withdrawn from an error model. Otherwise overfitting by too many parameters can occur and confidence intervals of estimated parameters can be increased, especially in the case of multicollinear effects (Markovitz *et al.*, 2005). For random effects, likelihood-ratio tests (Pinheiro and Bates 2000) and confidence intervals of estimated parameters are used to check significance.

Signal-to-noise ratios

$$\text{SNR} = \frac{\text{SD(predictions)}}{\text{SD(residuals)}} \tag{6}$$

allow a better interpretation of unexplained variability than AIC, BIC or P-values of ANOVA and are therefore calculated for the regarded error models, too. *Predictions* of a model are defined as the sum of all predictor variables.

Models leading to an accurate estimation of the dynamics are preferable. Therefore, standard errors $SE(T_{ost})$ should be small in comparison to estimated responses $\max_t(T_{ost}) - \min_t(T_{ost})$ after stimulation. For time-dependent proteins, we define a robust medial time response

$$\text{TR} = \text{median}_{os}\left(\frac{\max_t(T_{ost}) - \min_t(T_{ost})}{\text{median}_{\text{t}}(SE(T_{ost}))}\right). \tag{7}$$

Models leading to a small TR are not useful for the estimation of time courses.

Although a main purpose of our approach is the determination of time effects, time response TR is not used to select an appropriate model. Thereby, underestimation of confidence intervals by in sample optimization of the model is avoided. Nevertheless, time response constitutes an important property to validate the efficiency of a predetermined model.

In addition, a *leave one out cross-validation* (Efron, 1998) procedure is applied to determine the predictive power of the

models. Here, a model is iteratively fitted to all but one data points. This data point is then predicted by the fitted model. Accuracy of 'out-of-sample' predictions is measured via correlation of predicted and measured data and is used to evaluate the generalization error of the model. Under-parameterized or over-parameterized models would result in poor predictions.

A consistency criterion of an error model is that observed residuals correspond to assumed distribution of measurement errors. Differences between observed and theoretical distributions are assessed by a *Kolmogorov–Smirnov test* (Conover, 1971). On the one hand such a statistical test provide *P*-values p_{ks} to assess significance but, on the other hand, every error model will be rejected by a test asymptotically, e.g. for large sample size because theoretically assumed distribution will never be realized exactly. Nevertheless, the order of magnitude of *P*-values allows for comparisons between different models.

4 RESULTS

Benefits of log-transformation and multiplicative background correction are described in Section 4.1. In Section 4.2, models are fitted to housekeeping protein measurements to determine the most appropriate error model. The superior model is used to estimate time courses in Section 4.3 Additionally, it will be shown that model assessment criteria yield the same preferable model for time-dependent proteins. Figure 2 gives a schematic overview of applied analysis steps.

4.1 Background correction and response variables

For measurements of housekeeping proteins underlying protein concentrations are constant, e.g. independent on time and stimulation. Frequencies of measured intensities for housekeeping proteins are in agreement with lognormal distribution but disagree with normal distribution (Fig. 3). The parameters of normal and log-normal distribution are estimated by mean and variance of raw and log-transformed measurements.

A Kolmogorov–Smirnov test leads to $p_{ks} = 0.22$ for testing against log-normal distribution and rejects hypothesis of normally distributed noise with $p_{\mathrm{ks}} < 1E-19$. This result indicates that main sources of noise are multiplicative. In the Supplementary Material, comparison with normal and log-normal distribution is discussed in more detail.

Further, we find that intensity ratios

$$R = F/B \tag{8}$$

foreground over background are better reproducible than raw foreground intensities F or signals

$$S = F - B \tag{9}$$

obtained after background subtraction. Raw foreground intensities of repeated measurements show a coefficient of variation of around 40% within the same cell preparations and 27% within same preparations and on same gels. This variability is increased by background subtraction to around 57 and 38%, respectively. In contrast, ratios (8) reduce coefficients of variation to around 19 and 14%, for details see Supplementary material.

Biological, as well as technical noise seems to multiplicatively affecting foreground as well as background intensities. This influence is partly eliminated by calculation of intensity ratios. Emphasizing this hypothesis, foreground and background intensities are strongly correlated (see Supplementary Material). Beside raw foreground intensities and signals, we therefore consider additionally ratios R as well as $\log(S)$ and $\log(R)$ as response variables.

In addition to the six response variables shown in the upper part of Table 2, background can be considered by estimation of a regression parameter b between foreground and background intensities (lower part of Table 2). This proportionality constant can be fitted on intensity scale or on log scale. Special case $b = 1$ corresponds to background subtraction on intensity scale and coincides with intensity ratios for log-transformed intensities.

4.2 Error model selection for housekeeping proteins

Error models accounting for different systematic influences on measurements, e.g. observable effect O for different proteins, preparation effect P accounting for biological variability and gel effect G are fitted to our data using response variables of Table 2.

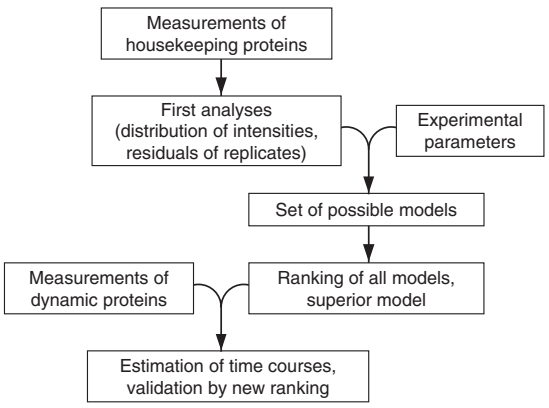

Fig. 2. Overview of applied analysis steps. First, analyses of housekeeping proteins enhance possibility of log-transformation and ratios foreground over background. In combination with different experimental parameters, 26 error models are introduced and evaluated. The superior model is used for determination of time courses.

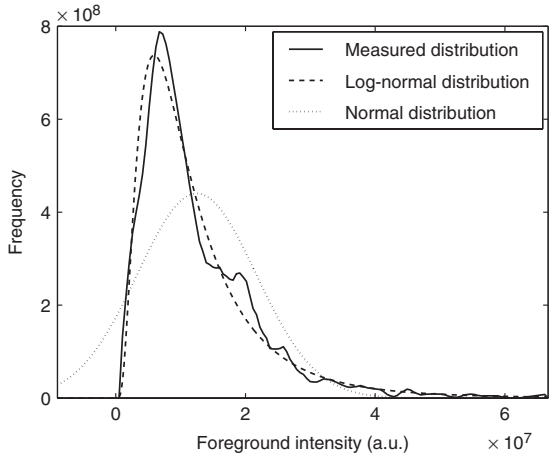

Fig. 3. Measured distribution of foreground intensities agrees with lognormal distribution. Its asymmetry is in contradiction with normal distribution.

Altogether, 26 combinations are considered as hypothetical error models. The different model assessment criteria introduced in Section 7.3 are displayed in Table 3 for all 26 models. The superior five values of each criterion are highlighted in bold face and the best values are underlined.

Akaike Information Criterion and Bayes' Information Criterion are superior for log(R) or log(F) with fitted log(B) (models 16–18 and 23–26). This advance does not strongly depend on the considered fixed and random effects. Additionally, residuals of these models are more consistent with normal distribution indicated by orders of magnitude larger P-values obtained by Kolmogorov–Smirnov tests.

Although both response variables show similar performance regarding model selection criteria and distribution of residuals, models with fitted background on log scale (models 23–26) are also superior concerning signal-to-noise ratio and cross validation. Therefore, log-transformation of measured intensities in combination with background correction on log–scale is recommended.

Additionally, there are strong indications that the impact of background differs from gel to gel. The best performance concerning Akaike and Bayes' Criteria as well as signal-to-noise ratio (6) shows model Number 26

$$\log(F_*) = O_o + \left(b + \epsilon_p^{(1)} + \epsilon_{pg}^{(1)}\right)\log(B_*) + \epsilon_p^{(2)} + \epsilon_{pg}^{(2)} + \varepsilon_* \ ,$$

$$\epsilon_p^{(1)} \sim N(0, \sigma_1^{(1)}), \epsilon_{pg}^{(1)} \sim N(0, \sigma_2^{(1)}), \epsilon_p^{(2)} \sim N(0, \sigma_1^{(2)}),$$

$$\epsilon_{pg}^{(2)} \sim N(0, \sigma_2^{(2)}), \varepsilon_* \sim N(0, \sigma), \qquad (10)$$

with random preparation and gel effects as well as a random preparation and gel-specific contribution to background correction. This model requires only 13 parameters because background dependency as well as preparation and gel effects are modeled as random effects, each with only a single parameter.

Table 2. Response variables of considered error models

Response y	Abbreviation	Background correction
Foreground	F	none
Signal	S	$F - B$
Signal ratio	R	F/B
log(foreground)	log (F)	none
log(signal)	log (S)	log ($F - B$)
log(signal ratio)	log (R)	log (F) – log (B)
Foreground, b fitted		F-b B
log(foreground), b fitted		log (F) – b log (B)

Response variables differ in log-transformation or in the way of accounting for measured background.

Table 3. Performance of error models for measurements of housekeeping proteins

Model Number	Model	n_{par}	AIC	n_{par}	p_{ks}	SNR	cor_{CV}
1	$F_* = O_o + \varepsilon_*$	6	39 100	39 100	5.4E-10	0.34	0.12
2	$F_* = O_o + P_p + G_{pg} + \varepsilon_*$	93	38 500	39 000	3.3E-10	1.1	0.68
3	$F_* = O_o + \varepsilon_p + \varepsilon_{pg} + \varepsilon_*$	8	38 500	38 500	1.1E-10	1.0	0.27
4	$S_* = O_o + \varepsilon_*$	6	38 000	38 100	3.7E-13	0.41	0.36
5	$S_* = O_o + P_p + G_{pg} + \varepsilon_*$	93	37 600	38 100	5.9E-8	1.0	0.61
6	$S_* = O_o + \varepsilon_p + \varepsilon_{pg} + \varepsilon_*$	8	37 500	37 600	3.2E-8	0.93	0.30
7	$R_* = O_o + \varepsilon_*$	6	2510	2540	3.7E-7	0.78	0.70
8	$R_* = O_o + P_p + G_{pg} + \varepsilon_*$	93	1990	2460	2.7E-6	1.4	0.79
9	$R_* = O_o + \varepsilon_p + \varepsilon_{pg} + \varepsilon_*$	8	2120	2160	1.3E-5	1.3	0.67
10	$\log(F_*) = O_o + \varepsilon_*$	6	2300	2330	**0.38**	0.31	0.13
11	$\log(F_*) = O_o + P_p + G_{pg} + \varepsilon_*$	93	1450	1920	0.016	1.3	0.78
12	$\log(F_*) = O_o + \varepsilon_p + \varepsilon_{pg} + \varepsilon_*$	8	1620	1660	**0.029**	1.2	0.34
13	$\log(S_*) = O_o + \varepsilon_*$	6	3120	3150	4.8E-4	0.36	0.35
14	$\log(S_*) = O_o + P_p + G_{pg} + \varepsilon_*$	93	2580	3050	6E-6	1.1	0.65
15	$\log(S_*) = O_o + \varepsilon_p + \varepsilon_{pg} + \varepsilon_*$	8	2720	2760	2.3E-5	0.98	0.40
16	$\log(R_*) = O_o + \varepsilon_*$	6	593	623	**0.84**	0.78	0.57
17	$\log(R_*) = O_o + P_p + G_{pg} + \varepsilon_*$	93	**−95**	**372**	0.016	1.6	0.78
18	$\log(R_*) = O_o + \varepsilon_p + \varepsilon_{pg} + \varepsilon_*$	8	**75**	**115**	0.023	1.5	0.43
19	$F_* = O_o + b\ B_* + \varepsilon_*$	7	37 600	37 600	7.8E-6	1.8	0.87
20	$F_* = O_o + b\ B_* + P_p + G_{pg} + \varepsilon_*$	94	37 200	37 600	9.5E-7	**2.5**	**0.95**
21	$F_* = O_o + b\ B_* + \varepsilon_p + \varepsilon_{pg} + \varepsilon_*$	9	37 100	37 200	3.1E-6	2.5	**0.89**
22	$F_* = O_o + (b + \varepsilon_p^{(1)} + \varepsilon_{pg}^{(1)})\ B_* + \varepsilon_p^{(2)} + \varepsilon_{pg}^{(2)} + \varepsilon_*$	13	36 900	36 900	7.0E-9	**3.0**	0.84
23	$\log(F_*) = O_o + b \log(B_*) + \varepsilon_*$	7	566	601	**0.96**	2.0	0.85
24	$\log(F_*) = O_o + b \log(B_*) + P_p + G_{pg} + \varepsilon_*$	94	**−93**	**379**	0.015	**3.1**	**0.95**
25	$\log(F_*) = O_o + b \log(B_*) + \varepsilon_p + \varepsilon_{pg} + \varepsilon_*$	9	**83**	**128**	**0.032**	**3.1**	**0.88**
26	$\log(F_*) = O_o + (b + \varepsilon_p^{(1)} + \varepsilon_{pg}^{(1)}) \log(B_*) + \varepsilon_p^{(2)} + \varepsilon_{pg}^{(2)} + \varepsilon_*$	13	**−98**	**−33**	0.008	**3.7**	**0.90**

Abbreviation '$*$' is used instead of all occurring indices in a model, e.g. indices of all predictor variables and an index for replicate measurements. The superior five values of each model assessment criterion are highlighted in bold face, best values are underlined. A log-transformation improves the performance of the models. Model 26 is superior for 3 out of 5 criteria.

Table 4. Model assessment criteria for time-dependent proteins

Model	Response variable	AIC	BIC	p_{ks}	SNR	TR
4′	Signals	55 345	55 740	3.4*E*-83	0.3	2.4
16′	Log-ratios	807	1202	0.031	0.6	5.1
6′	Model for signals	51 691	52 096	9.5*E*-39	1.6	5.8
26′	Complete model	490	923	0.0052	4.2	8.0

Error models for housekeeping proteins are extended by time effects yielding to models 1′ to 26′. Here, the performance of a subset of four prominent models is displayed for the full data set. Again, the complete error model shows preferable model assessment criteria. Complete table is provided in Supplementary Material.

ANOVA of fixed effects leads to significant *P*-values. Additionally, 95% confidence intervals of estimated random effects are significantly different from zero. Therefore, impact of preparation and gel effects should be accounted if immunoblotting is used quantitatively. This tendency is obtained for all regarded response variables.

In addition to discussed experimental parameters, no improvement could be observed by effects accounting for the size of spots or for lane-specific differences within gels. Spot size effects show strong correlations with underlying concentrations and lead therefore for inflating standard errors of estimated protein concentrations. Lane effects are partly accounted by locally determined background and are only badly identifiable due to limited number of observables within a lane.

For our analyses, we assume exclusively additive or multiplicative noise. The case of both, additive and multiplicative noise is discussed in Supplementary Material.

4.3 Application to time course measurements

To illustrate the benefits of log-transformation and elimination of systematic noise, error models are now applied to determine time courses of insulin signaling. Therefore, models for housekeeping proteins are enlarged by time effects T_{ost} which depend on observable o and applied stimulation s. Model (10) which was superior for housekeeping proteins yields

$$\log(F_*) = O_o + T_{ost} + \left(b + \epsilon_p^{(1)} + \epsilon_{pg}^{(1)}\right)\log(B_*) + \epsilon_p^{(2)} + \epsilon_{pg}^{(2)} + \varepsilon_*. \tag{11}$$

Observable effects O_o are chosen in a way that $T_{ost_0} = 0$ is fulfilled. A table with model assessment criteria of all 26 models for time-dependent proteins is provided in Supplementary Material. Affirming the results obtained for housekeeping proteins, model 26, respeatively. Equation (11), remains the most appropriate model for time-and stimulation-dependent proteins. It has superior model selection criteria, shows almost normally distributed residuals and a preferable signal-to-noise ratio.

The improvement of model assessment criteria by log-transformation and multiplicative background correction is illustrated in Table 4. Here, model 26′ is compared with three prominent out of 25 competing other models. In first, model 4′, time effects are estimated from signal intensities without considering preparation or gel effects. Log-ratios are used for the second model Number 16′. The third selected model 6′ corresponds to model 26′ but no log-transformation is applied.

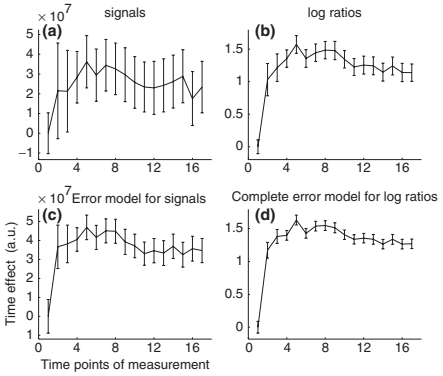

Fig. 4. Activation of Insulin receptor estimated (**a**) from signals, (**b**) from log-ratios, (**c**) from signals with elimination of systematic errors and (**d**) for the full model with log-transformation and multiplicative background correction.

Model assessment criteria are clearly improved by the full model Number 26′.

The benefit of model 26′ results in a better estimation of time effects. As an example, estimation of time dependency of phosphorylated IR* after stimulation with 100 nM insulin and standard errors are plotted in Figure 4. Using the complete model (Fig. 4 (d)) leads to reliable estimation of time dependency and small confidence intervals. In contrast to this, without log-transformation or if systematic errors are not regarded (Fig. 4a–c) signal-to-noise ratio and smoothness of estimated time courses are decreased. Figures for all 26 models are displayed in Supplementary Material.

5 CONCLUSION

Molecular biology has already made remarkable contribution to our understanding of biological systems on a cellular level. However, in recent years it became obvious that screening of all molecular components is not enough to understand complexity of cellular processes. Time resolved measurements in combination with mathematical models have to be used to uncover dynamics and systems properties of biochemical networks.

Particularly in Systems Biology, integration of both experimental data and mathematical modeling is used to discover principles of cell biology. Systems Biology approaches are successfully applied to study signaling pathways in cells (Bentele *et al.*, 2004; Schoeberl *et al.*, 2002).

Here, we used time resolved data of insulin signaling pathway. Development and validation of a mathematical model for dynamics of protein concentrations requires not only qualitative but also quantitative reproducible measurements. This requirement could only be achieved by development of an error model for immunoblotting intensities.

Despite development of new experimental techniques, quantitative reliability constitutes a main bottleneck of new experimental approaches. Many methods show a minor signal-to-noise ratio and are only limited applicable for time resolved measurements, especially if primary cells are studied. Statistical analysis of measurement noise by error models has been successfully applied to some other experimental techniques, e.g. for

high-throughput screening techniques (Malo *et al*., 2006), analytical chemistry (Rocke and Lorenzato, 1995), microarrays (Love *et al*., 2002; Weng *et al*., 2006) or polymerase chain reactions Pfaffl *et al*., 2001). A challenging next step is the statistical analysis of data obtained by other experimental techniques.

We introduced and compared error models for data obtained by immunoblotting. It could be shown by a model selection procedure that predominant errors are multiplicative. A log-transformation of measured intensities is required to obtain desired statistical properties like Gaussian distribution of noise. These statistical properties are necessary for dynamic modeling of protein interactions, e.g. in signal transduction. A consequence is that error bars for immunoblotting intensities are not symmetric. Although outliers towards large values are often observed, asymmetry of error bars are not considered in practice so far.

We recommend that background correction procedures are applied on log scale. Furthermore, a gel-specific background correction seems preferable.

It is strongly recommended to determine background intensities carefully. Overestimation of local background by contamination with signals would lead to a lower signal-to-noise ratio after correction. To avoid this risk, we checked that the same results are obtained by using background intensities determined 'far away' from signal spots on our gels.

Furthermore, we revealed strong biological variability between cell preparations as well as technical noise between different gels. Elimination of these sources of a bias improves reproducibility of the data significantly resulting in smaller error-bars of protein concentration time courses.

In comparison to separate calibration experiments (Schilling *et al*., 2005b), our approach does not require additional experiments. The amount of observational noise is determined by replicate measurements from time course measurement itself. The advantage of this approach is that calibration experiments can be avoided and the error model is developed in exactly the same experimental setup as time courses, e.g. same cells, proteins and antibodies. Nevertheless, calibration experiments can be used to estimate the fraction of systematic errors which are assumed to be equal in calibration and time course data.

We found a coefficient of variation within gels of around 14%. For simulation studies generating synthetic data of protein concentrations, a realistic setting for immunoblotting data would be obtained by generating log-normally distributed noise with a similar coefficient of variation.

Modeling of measurements and measurement errors allows for statistical testing and elimination of systematic errors. The illustrated mixed effects model approach can be extended to correct for arbitrary undesired influences, e.g. by different experimental techniques, different cell types or tissues. Therefore, mixed effects models constitute an appropriate approach to perform an analysis of data obtained under various experimental conditions. This integration of different data sources is an important and essential step for systems biology.

After all, development of appropriate error models accounting for biological inhomogeneity and experimental noise is one step in the key challenge of systems biology, the generation of reliable and biologically relevant mathematical models.

ACKNOWLEDGEMENTS

The authors thank Ursula Klingmueller, Jan Hengstler and their groups for experienced support. In addition, the authors gratefully acknowledge financial support by the grant BMBF 0313074D & FP6 EU-grant COSBICS LSHG-CT-2004-0512060.

Conflict of Interest: none declared.

REFERENCES

Atkinson,A. (1981) Likelihood ratios, posterior odds and information criteria. *J. Econom.*, **16**, 15–20.

Bentele,M. *et al*. (2004) Mathematical modeling reveals threshold behavoiur of CD95-induced apoptosis. *J. Cell Biol.*, **166**, 839–851.

Bolstad,B. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Conover,W. (1971) *Practical Nonparametric Statistics*. John Wiley & Sons, New York.

Durbin,B. and Rocke,D. (2003) Estimation of transformation parameters for microarray data. *Bioinformatics*, **20**, 660–667.

Efron,B. and Tibshirani,R.J. (1998) *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, 96–104.

Ideker,T. *et al*. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.

Kurien,B. and Scofield,R. (2006) Western blotting. *Methods*, **38**, 283–293.

Limpert,E., Stahel,W. and Abbt,M. (2001) Log-normal distributions across the sciences: keys and clues. *Bioscience*, **51**, 341–352.

Love,B. *et al*. (2002) A conditional density error model for the statistical analysis of microarray data. *Bioinformatics*, **18**, 1064–1072.

Malo,N. *et al*. (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.*, **24**, 167–175.

Markovitz,B. (2005) The principle of multicollinearity. *Pediatr. Crit. Care Med.*, **6**, author reply 94–95.

Pavelka,N. *et al*. (2004) A power law global error model for the identification of differentially expressed genes in microarray data. *BMC Bioinformatics*, **5**, 203.

Pfaffl,M. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.*, **29**, 45.

Pinheiro,C. and Bates,M. (2000) *Mixed-Effects Models in S and S-Plus*. Springer, New York.

Rocke,D. and Durbin,B. (2003) Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, **19**, 966–972.

Rocke,D. and Lorenzato,S. (1995) A two-component model for measurement error in analytical chemistry. *Technometrics*, **37**, 176–185.

Sakamoto,Y. *et al*. (1986) *Akaike Information Criterion Statistics*., D. Reidel Publishing Company.

Schilling,M. *et al*. (2005a) Computational processing and error reduction strategies for standardized quantitative data in biological networks. *FEBS J.*, **272**, 6400–6411.

Schilling,M. *et al*. (2005b) Quantitative data generation for Systems Biology: the impact of randomisation, calibrators and normalisers. *IEE Proc. Syst. Biol.*, **152**, 193–200.

Schoeberl,B. *et al*. (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol*, **20**, 370–375.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Swameye,I. *et al*. (2003) Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling. *Proc. Natl. Acad. Sci.* USA, **100**, 1028–1033.

Weng,L. *et al*. (2006) Rosetta error model for gene expression analysis. *Bioinformatics*, **22**, 1111–1121.

Yang,Y.H. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.