Frank Klawonn

Significance tests to identify regulated proteins based on a large number of small samples

Kybernetika, Vol. 48 (2012), No. 3, 478--493

Persistent URL: `http://dml.cz/dmlcz/142950`

KYBERNETIKA — VOLUME *48* (2012), NUMBER 3, PAGES 478–493

SIGNIFICANCE TESTS TO IDENTIFY REGULATED PROTEINS BASED ON A LARGE NUMBER OF SMALL SAMPLES

Frank Klawonn

Modern biology is interested in better understanding mechanisms within cells. For this purpose, products of cells like metabolites, peptides, proteins or mRNA are measured and compared under different conditions, for instance healthy cells vs. infected cells. Such experiments usually yield regulation or expression values – the abundance or absence of a cell product in one condition compared to another one – for a large number of cell products, but with only a few replicates. In order to distinguish random fluctuations and noise from true regulations, suitable significance tests are needed. Here we propose a simple model which is based on the assumption that the regulation factors follow normal distributions with different expected values, but with the same standard deviation. Before suitable significance tests can be derived from this model, a reliable estimation for the standard deviation in the context of many small samples is needed. We therefore also include a discussion on the properties of the sample MAD (**M**edian **A**bsolute **D**eviation from the median) and the sample standard deviation for small samples sizes.

Keywords: MAD, standard deviation, small samples, significance test

Classification: 93E12, 62A10

1. INTRODUCTION AND MOTIVATION

In order to better understand the mechanisms within cells, biologists analyse the genome, the proteome and the metabolome[1]. Typical experiments compare different conditions or states like healthy cells with infected or cancerous cells, bacteria with and without the presence of antibiotics or cells at different time points after a certain stimulation. Modern high-throughput technologies like mass spectrometry for proteomics and metabolomics or microarrays and next generation sequencing for genomics enable biologists to compare such conditions in terms of regulation factors or expression values. Such experiments provide data to quantify the abundance or absence of proteins, metabolites or gene products (for instance, in terms of mRNA) in one condition compared to another one. The resulting data are often regulation factors or expression values specifying, for example, that in condition B (e. g. cancerous cell) there was three times as much of protein prot_i than in condition A (e. g. healthy cell).

[1] The genome, the proteome and the metabolome represent the sets of all genes, proteins respectively metabolites that can be produced or expressed within a cell

All these experiments have in common that a larger number of "items" (gene products, proteins, peptides, metabolites) – from a few hundred up to tens of thousands – are measured, but only few replicates are available, normally a number between 2 and 10. High costs for the experiments, a limited number of available donors – e. g. specific cancer patients – or ethical considerations for animal experiments are the reasons for the small number of replicates.

Uncertainty in measurement caused by the measurement technique and an often high variation of the biological system is very common for these experiments. Therefore, it is critical to distinguish random fluctuations from relevant or statistically significant regulations. There are a variety of special data analysis models for each of these measurement techniques: For microarrays see for instance [3, 6, 15] for next generation sequencing [1, 10, 12] and for mass spectrometry [4, 8, 9].

Although we focus on the analysis of proteomics mass spectrometry data in this paper, our approach can in principle also be applied to other types of regulation or expression data. Of course, methods that are specifically tailored to devices like [4, 8, 15] can lead to more precise results. However, these special methods have also their disadvantages.

- They often require additional experiments for calibrations and estimation of model parameters for the specific device or experimental setting [8].
- They are more difficult to understand and to apply and require more computational costs. LIMMA [15], a very common approach to analyse microarray data is based on Bayesian methods and is only applicable in the context of microarray experiments.
- They usually operate on raw data or on data after simple preprocessing steps. For published data that might be of interest for an analysis from a different perspective, the raw data are often not available and the model parameters for the device are unknown.

Furthermore, the method proposed here is quite general and generic and therefore comparisons between experiments – even if they were carried out on different devices – can be performed easily. We do not claim that our method is better in general. It is especially suitable for proteomics experiments, but could in principle also be applied in other contexts, although – especially for microarrays – it might be more recommendable to apply specifically tailored techniques like LIMMA [15].

The remainder of the paper is organised as follows. In the next section, the data analysis problem is presented in a more formal way. Section 3 describes the general approach to estimate the variance in the data which will be needed for defining statistical tests for significant regulation later on. Section 4 reviews some properties of measures of spread that can be used to estimate the standard deviation of our model distribution. Two significance tests for regulation are then proposed in Section 5 and are evaluated empirically in Section 6 before the final concluding remarks.

2. PROBLEM FORMULATION

In this section, we describe the problem in a more formal way. The data sets, we have to deal with, usually have the form as shown in Table 1. As mentioned before, the number

Protein	Replicate R_1	...	Replicate R_n
Prot_1	$r_{1,1}$	...	$r_{1,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Prot_k	$r_{k,1}$	...	$r_{k,n}$

Tab. 1. Structure of a typical data table of regulation factors.

n of replicates is usually very small, often in the range between 2 and 10, whereas the number n of proteins is often a few thousand. We only consider two conditions at a time, so that the entries in the table correspond to the log2-regulation factors between the first condition A and the second condition B. If the measured intensity for protein Prot_i in condition A was 1000 and in condition B 2000 in replicate R_j, then we would have $r_{i,j} = 1 = \log_2\left(\frac{2000}{1000}\right)$. The logarithmic regulation is considered in order to treat up-regulation – a higher value for a protein in condition B than in condition A – and down-regulation – a lower value for a protein in condition B than in condition A – in a symmetric manner. Otherwise, without the logarithm up-regulation would be in the range $(1, \infty)$, whereas down-regulation would be squeezed into the interval $(0, 1)$.

We restrict our considerations to the comparison between two conditions. If there are more than two conditions – for instance more than two time points or different kinds of infections – our methods can be applied to each pair of conditions separately.

It should be noted that we usually cannot avoid that the table contains a certain fraction of missing values, since sometimes single proteins cannot be identified or measured properly in one or some of the replicates.

The goal is to distinguish significant regulations of a protein from random fluctuations. If we consider the information that we have from the few replicates for each protein separately, it is more or less impossible to identify significant deviations from 0, i. e. from no regulation. We would need a reliable estimation of the random variation for each protein which cannot be obtained from a very small sample. However, we can assume a certain common characteristic for all proteins and in this way derive more reliable estimations for the variance and deviations from no regulation. The underlying model will be discussed in the following section.

3. MODEL ASSUMPTIONS

In order to identify protein regulations that are not due to random fluctuations, it is necessary to estimate the noise or variation of regulation factors for each protein. Due to the very limited number of replicates, it is impossible to provide a reliable estimation of the variation for each protein individually. However, based on the assumptions

- that the (random) variation of the regulation factors within replicates for a protein is independent of the (true) regulation factor and
- that the regulation factors of the replicates follow a normal distribution,

a very reliable estimation of the standard deviation of these normal distributions can be provided.

From a statistical point of view, we have for each a protein a small sample of regulation factors. The sample size for each protein corresponds to the number of replicates n, at least if there are no missing values. Under the above assumptions, all these small samples originate from normal distributions with different expected values (the "true" regulation factors), but with the same standard deviation.

Therefore, for each protein we can obtain an unreliable estimation of this standard deviation. Since the number of (observed) proteins is quite large, we obtain a large number of unreliable estimations for the same standard deviation. Aggregating these unreliable estimations together (by the mean value, median or trimmed mean), we obtain a reliable estimation of the standard deviation.

Our main interest is specifically the characterisation of the distribution of measured regulation factors of unregulated proteins. This means that we are primarily interested in the standard deviation of the (assumed) normal distribution with expected value 0, representing no regulation. Our assumption that the standard deviation is independent of the underlying "true" regulation factor might not be correct. Typically, the regulation factors of proteins with higher regulations also show a higher variance. Since we will use the single estimates of the standard deviation for all proteins, this means that we tend to overestimate the standard deviation at 0. Therefore, the tests that we will develop and which rely on the estimation of the standard deviation of the normal distribution at 0 will tend to be more conservative in the sense that we might not find all regulated proteins.

Figure 1 illustrates our simple model and the principle of obtaining a reliable estimation of the standard deviation from a larger number of small samples. There are $n = 3$ replicates and $k = 4$ proteins in Figure 1. The number of replicates is realistic although it might sometimes be a little bit higher. The number of proteins is kept small for illustration purposes. In reality, this number can easily exceed 1000. An estimation of a standard deviation based on a sample of size $n = 3$ is not reliable at all. Any estimator of this standard deviation will have a high variance. If we, however, combine all these unreliable estimators, for instance by taking the mean, this new estimator will have a much smaller variance. The variance of this estimator would be decreased by the factor $1/\sqrt{k}$. Since k will be a larger number in our case, we do not need to worry too much about the variance of the unreliable estimator. But we require an unbiased estimator. Only in this way, we can guarantee consistency of the aggregated estimator where consistency here refers to $k \to \infty$.

We will discuss two estimators for the standard deviation of the normal distribution here: the sample standard deviation and the sample MAD (**M**edian **A**bsolute **D**eviation from the median). The sample standard deviation is a consistent, but biased estimator for the standard deviation. And the sample MAD is a consistent, but biased estimator for the MAD and can be turned into an estimator for the standard deviation by multiplying it with a suitable constant. Since we have to deal with small samples sizes – although we will aggregate many of them – a correction is necessary to obtain unbiased estimators. Note that an increase in the number k of proteins will not contribute to a correction of the bias of these estimators.

We restrict our discussion here to the standard deviation as an efficient estimator in the context of a normal distribution and the MAD as a very robust estimator, especially

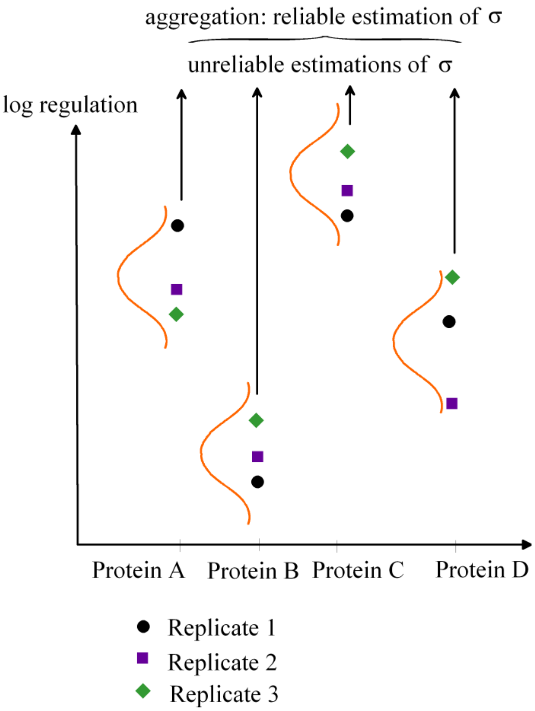

Fig. 1. Principle of the estimation of the standard deviation σ with a (larger) number of small samples.

since our assumption of normal distribution might be violated for larger regulation factors. Of course, there are also alternatives that are discussed in detail as estimators for the standard deviation in [13] based on a sample $x_1, \ldots, x_n$. These estimators are

$$S_n^{(\mathrm{RC})} = a_n \cdot \mathrm{med}_i \{\mathrm{med}_j |x_i - x_j|\}$$

and the ℓth order statistic of the absolute differences between the sample values i. e.

$$Q_n = b_n \cdot \{|x_i - x_j|; i < j\}_{(\ell)}$$

where $\ell = \binom{[n/2]+1}{2}$. In both cases, a suitable constant a_n respectively b_n is introduced to obtain an unbiased estimator for the standard deviation. A detailed discussion of these estimators for our purposes would be out of the scope of this paper.

In the following section we take a closer look at the bias of the standard deviation and the MAD for normal distributions.

4. MAD VS. STANDARD DEVIATION

At least for continuous distributions, sample quantiles are consistent estimators for the true quantiles and therefore the sample MAD is also a consistent estimator for the true

MAD. For a normal distribution we have

$$\sigma = \Phi^{-1}\left(\frac{3}{4}\right) \cdot \mathrm{MAD} \;\approx\; 1.4826 \cdot \mathrm{MAD} \tag{1}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. However, because the sample MAD is a biased estimator, the constant $\Phi^{-1}\left(\frac{3}{4}\right)$ in Eq. (1) can only be used for large sample sizes. Since we have to deal with small samples, although many of them, we cannot simply use this constant.

We denote the sample MAD and the sample standard deviation for a sample of size n by MAD_n and S_n, respectively.

We now take a closer look at the sample MAD and the sample standard deviation for normal distributions, first on a purely theoretical basis for the very small sample size $n = 2$ and then for $n > 2$ based on a simulation study.

4.1. The case $n = 2$

In order to examine the properties of MAD_2 and S_2, we need the following lemmata.

Lemma 4.1. Let $Y \sim N(0, \sigma^2)$ where $N(\mu, \sigma^2)$ denotes a normal distribution with expected value μ and variance σ^2. Then the following equations hold.

(a) $\mathrm{E}(|Y|) = \sqrt{\frac{2}{\pi}} \cdot \sigma$.

(b) $\mathrm{E}(|Y|^2) = \sigma^2$.

(c) $\mathrm{Var}|Y| = \frac{\pi-2}{\pi} \cdot \sigma^2$.

Proof.

(a) $\mathrm{E}(|Y|) = \dfrac{1}{\sqrt{2\pi} \cdot \sigma} \displaystyle\int_{-\infty}^{\infty} e^{-\dfrac{x^2}{2\sigma^2}} \cdot |x| \, \mathrm{d}x = \dfrac{2}{\sqrt{2\pi} \cdot \sigma} \displaystyle\int_{0}^{\infty} e^{-\dfrac{x^2}{2\sigma^2}} \cdot x \, \mathrm{d}x = \sqrt{\dfrac{2}{\pi}} \cdot \sigma.$

(b) $\mathrm{E}(|Y|^2) = \mathrm{E}(Y^2) = \mathrm{Var}(Y) - (E(Y))^2 = \sigma^2$.

(c) Using (a) and (b), we obtain

$$\mathrm{Var}|Y| = E(|Y|^2) - (E(|Y|))^2 = \frac{\pi - 2}{\pi} \cdot \sigma^2.$$

□

Lemma 4.2. Let $X_1, X_2 \sim N(\mu, \sigma^2)$ i.i.d. Then

$$\frac{X_1 - X_2}{2} \;\sim\; N\left(0, \frac{\sigma^2}{2}\right)$$

holds.

Proof. The sum of independent normal distributions follows again a normal distribution. Also the multiplication with a constant factor will lead to a normal distribution. Therefore, $\frac{X_1-X_2}{2}$ follows a normal distribution.

$$\mathrm{E}\left(\frac{X_1-X_2}{2}\right)=\frac{1}{2}\cdot(\mathrm{E}(X_1)-E(X_2))=0.$$

$$\mathrm{Var}\left(\frac{X_1-X_2}{2}\right)=\frac{1}{4}\cdot(\mathrm{Var}(X_1)+\mathrm{Var}(X_2))=\frac{\sigma^2}{2}.$$

□

Proposition 4.3. Let $X_1, X_2 \sim N(\mu, \sigma_0^2)$ i.i.d. Then the following equations hold.

(a) $\mathrm{E}\left(\mathrm{MAD}_2\right)=\frac{\sigma_0}{\sqrt{\pi}}$.

(b) $\mathrm{Var}\left(\mathrm{MAD}_2\right)=\frac{\pi-2}{2\cdot\pi}\cdot\sigma_0^2$.

Proof.

(a) The median of X_1 and X_2 is in this case the same as the mean $\frac{X_1+X_2}{2}$. The absolute deviation of X_1 and X_2 from the median is in both cases

$$\left|X_1-\frac{X_1+X_2}{2}\right|=\left|X_2-\frac{X_1+X_2}{2}\right|=\left|\frac{X_1-X_2}{2}\right|.$$

From Lemma 4.2 we know that $\frac{X_1-X_2}{2}\sim N\left(0,\frac{\sigma_0^2}{2}\right)$ holds. Applying Lemma 4.1(a) yields $\mathrm{E}\left(\mathrm{MAD}_2\right)=\frac{\sigma_0}{\sqrt{\pi}}$.

(b) According to Lemma 4.2 we have $\frac{X_1-X_2}{2}\sim N\left(0,\frac{\sigma_0^2}{2}\right)$. From Lemma 4.1(c) we obtain immediately $\mathrm{Var}\left(\mathrm{MAD}_2\right)=\frac{\pi-2}{2\cdot\pi}\cdot\sigma_0^2$.

□

Since we know from Eq. (1) that $\mathrm{MAD}=\frac{\sigma_0}{\Phi^{-1}\left(\frac{3}{4}\right)}\approx\frac{\sigma_0}{1.4826}$ holds for the true MAD, Proposition 4.3 shows that MAD_2 is a biased estimator for the true MAD of a normal distribution.

There is a simple connection between MAD_2 and S_2 as the following proposition shows.

Proposition 4.4. Let $X_1, X_2 \sim N(\mu, \sigma_0^2)$ i.i.d. Then

$$\mathrm{S}_2=\sqrt{2}\cdot\mathrm{MAD}_2$$

holds.

Proof. It is obvious that $\mathrm{S}_2^2=2\cdot\mathrm{MAD}_2^2$ holds. □

The following corollary shows that S_2 is a biased estimator for the standard deviation of a normal distribution.

Corollary 4.5. Let $X_1, X_2 \sim N(\mu, \sigma_0^2)$ i.i.d. Then the following equations hold.

(a) $\mathrm{E}\left(\mathrm{S}_2\right) = \sqrt{\frac{2}{\pi}} \cdot \sigma_0$.

(b) $\mathrm{Var}\left(\mathrm{S}_2\right) = \frac{\pi-2}{\pi} \cdot \sigma_0^2$.

Proof. This corollary is an immediate consequence of Propositions 4.4 and 4.3. □

We can now turn MAD_2 and S_2 into unbiased estimators for the standard deviation of a normal distribution by multiplying them with suitable constants. Both estimators have the same efficiency.

Corollary 4.6. Let $X_1, X_2 \sim N(\mu, \sigma_0^2)$ i.i.d. Then

(a) $\mathrm{MAD}_2^{(\mathrm{cor})} = \sqrt{\pi} \cdot \mathrm{MAD}_2$ and

(b) $\mathrm{S}_2^{(\mathrm{cor})} = \sqrt{\frac{\pi}{2}} \cdot \mathrm{S}_2$

are unbiased estimators for σ_0. Both estimators have the variance

$$\mathrm{Var}\left(\mathrm{MAD}_2^{(\mathrm{cor})}\right) = \mathrm{Var}\left(\mathrm{S}_2^{(\mathrm{cor})}\right) = \frac{\pi - 2}{2} \cdot \sigma_0^2.$$

Proof. This corollary is an immediate consequence of Proposition 4.3 and Corollary 4.5. □

It should be noted that even the sample standard deviation needs a correction of more than 25 % for this extremely small sample size in order to become an unbiased estimator.

Unfortunately, these theoretical considerations cannot be easily extended to $n > 2$. Therefore, we will consider sample sizes larger than two based on simulation studies.

4.2. Arbitrary sample sizes

In order to determine the correction factors to turn MAD_n into an unbiased estimator for $n > 2$ in our context, we will carry out a simulation study. Croux and Rousseeuw [5] have already carried out a similar simulation study on the bias of MAD_n and the estimators[2] S_n^{RC} and Q_n mentioned in the previous section. The differences between the simulation study of Croux and Rousseuw and ours are that we include the sample standard deviation and that we take a slightly different view on the variance of the estimators. The latter difference arises from the fact that we consider a larger number of small samples.

Table 2 shows the result of our simulation study based on a standard normal distribution. We have chosen $k = 1000$, so that we obtain 1000 different estimations for MAD_n and S_n for different values of n. We then take the mean value of these 1000 estimations. We repeat this procedure 100 times, so that we finally have 100 estimations of MAD_n

n	MAD	MAD variance	SD	SD variance
2	0.5636	1.65E-05	0.7978	3.68E-05
3	0.4535	1.33E-05	0.8868	2.30E-05
4	0.4959	7.17E-06	0.9210	1.74E-05
5	0.5538	1.09E-05	0.9400	1.13E-05
6	0.5673	7.66E-06	0.9514	1.04E-05
7	0.5925	8.82E-06	0.9593	8.09E-06
8	0.5982	4.81E-06	0.9653	5.82E-06
9	0.6125	6.15E-06	0.9692	4.92E-06
10	0.6152	4.27E-06	0.9728	4.88E-06
20	0.6471	2.75E-06	0.9871	2.67E-06
30	0.6567	1.90E-06	0.9914	1.39E-06
40	0.6610	1.37E-06	0.9935	1.39E-06
60	0.6659	9.16E-07	0.9957	8.01E-07
80	0.6680	6.76E-07	0.9968	7.33E-07
100	0.6692	4.52E-07	0.9975	4.87E-07
120	0.6701	4.65E-07	0.9980	4.70E-07
150	0.6710	4.33E-07	0.9983	3.09E-07
200	0.6719	2.71E-07	0.9988	1.66E-07
300	0.6727	2.30E-07	0.9991	1.57E-07
500	0.6735	1.03E-07	0.9995	8.71E-08
1000	0.6740	6.22E-08	0.9997	4.84E-08
∞	0.6745	0	1	0

Tab. 2. Estimated correction factors and their variance.

and S_n in the setting of n replicates and $k = 1000$ proteins. The sample variance of these 100 estimates is also included in the table.

Figure 2 illustrates the bias of MAD_n and S_n for $n \in \{2, \ldots, 10\}$. The horizontal lines indicate the true values for the MAD and the standard deviation, respectively. It is remarkable that the bias for MAD_2 is smaller than for MAD_3, MAD_4 and MAD_5. This figure also shows that the bias of MAD_n decreases slower than the bias of S_n.

Figure 2 just magnifies the range for smaller sample sizes n which are of special interest for us. Figure 3 extends Figure 2 to $n = 100$.

Table 3 is directly derived from Table 2 and shows the correction factors to turn MAD_n and S_n into unbiased estimators for the standard deviation of a normal distribution. Note that the correction factor for the sample size $n = 3$ is almost 50 % larger than the correction for large sample sizes for MAD_n. Even for S_3 a correction of more than 10 % is needed. The table shows also that S_n is a more efficient estimator than MAD_n in the case of a normal distribution. Nevertheless, we prefer to use MAD_n in our context, since

[2] Note the difference between the sample standard deviation S_n and the estimator $S_n^{(\mathrm{RC})}$. Croux and Rousseuw use the notation S_n instead of $S_n^{(\mathrm{RC})}$ in their study. We deviate from this notation in order to avoid confusion with the sample standard deviation.

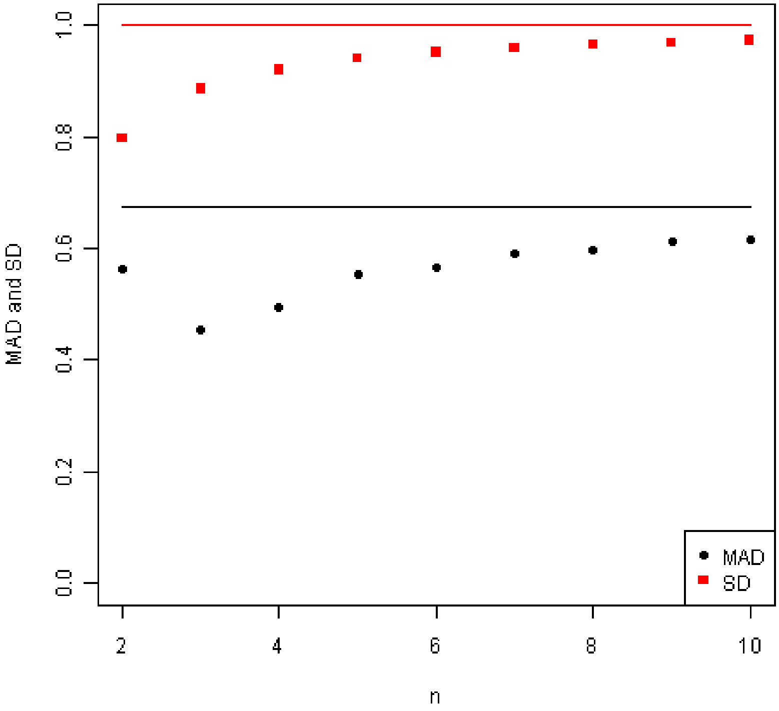

Fig. 2. MAD and standard deviation estimators depending on the sample size.

our assumption of normal distribution might be violated for larger regulation factors and can lead to extreme outliers for S_n for small n.

5. COMPUTATION OF p-VALUES FOR SIGNIFICANT REGULATION

Based on the assumption that the regulation factors for each protein follow normal distributions with the same standard deviation σ, but different expected values, we can construct a hypothesis test for the identification of proteins with significant regulation, for which we use the estimate σ_0 computed as described in the previous section. The null hypothesis for this hypothesis test will be that the considered protein is not regulated, so that its log2-regulation factors are assumed to follow a normal distribution with expected value 0 and variance σ_0^2. Note that for this null hypothesis we actually only require that the regulation factors of unregulated proteins follow a normal distribution and that we have a good or at least a conservative estimate of the standard deviation given the null hypothesis. In the following, we propose two tests for the identification of significantly

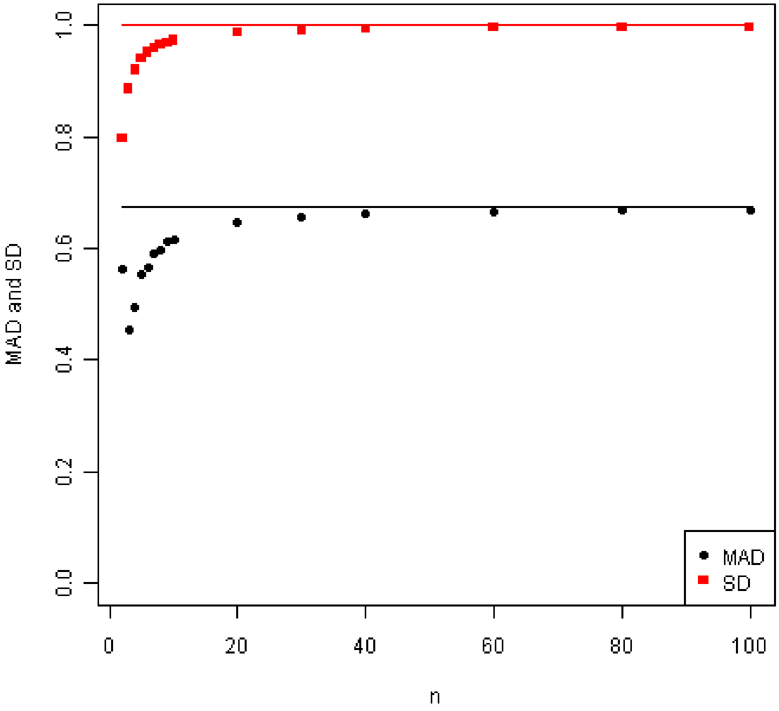

Fig. 3. MAD and standard deviation estimators depending on the sample size.

regulated proteins.

5.1. p-values based on a simple Z-test

When we focus on a single protein prot$_i$ and are only interested whether this specific protein shows a significant regulation, we can apply a simple Z-test, i. e. a simplified Student's t-test with known variance where we use the estimate σ_0^2 for the variance from the previous section. Under the null hypothesis that the protein is not regulated and thus its log2-regulation factors should follow a normal distribution with expected value 0 and variance σ_0^2, the test statistic

$$T = \frac{\bar{X}}{\sigma_0/\sqrt{\tilde{k}}}$$

n	MAD factor	MAD variance	SD factor	SD variance
2	1.7744	5.19E-05	1.2535	5.79E-05
3	2.2053	6.48E-05	1.1276	2.92E-05
4	2.0164	2.91E-05	1.0857	2.05E-05
5	1.8056	3.55E-05	1.0638	1.28E-05
6	1.7628	2.38E-05	1.0511	1.15E-05
7	1.6876	2.51E-05	1.0424	8.79E-06
8	1.6717	1.34E-05	1.0359	6.25E-06
9	1.6326	1.64E-05	1.0318	5.24E-06
10	1.6254	1.13E-05	1.0280	5.15E-06
20	1.5454	6.57E-06	1.0130	2.74E-06
30	1.5228	4.41E-06	1.0087	1.41E-06
40	1.5129	3.14E-06	1.0065	1.41E-06
60	1.5017	2.07E-06	1.0043	8.08E-07
80	1.4970	1.51E-06	1.0032	7.38E-07
100	1.4942	1.01E-06	1.0026	4.89E-07
120	1.4922	1.03E-06	1.0020	4.72E-07
150	1.4903	9.62E-07	1.0017	3.10E-07
200	1.4884	6.00E-07	1.0012	1.67E-07
300	1.4864	5.07E-07	1.0009	1.58E-07
500	1.4849	2.27E-07	1.0005	8.72E-08
1000	1.4836	1.37E-07	1.0003	4.84E-08
∞	1.4826	0	1	0

Tab. 3. Estimated correction factors and their variance.

yields a standard normal distribution with expected value 0 and variance 1. $\bar{X}$ is the mean value of the measured log2-regulation factors of protein prot$_i$, and $\tilde{k} \leq k$ is the number of replicates in which log2-regulation factors have been measured for protein prot$_i$. Therefore, the p-value for a two-sided test with the alternative hypothesis that the log2-regulation factors of the considered protein do not follow a normal distribution with expected value 0 can be computed as follows.

$$\begin{aligned} p_i &= P(T > t \vee T < -t) \\ &= P(|T| > t) \\ &= 2 \cdot (1 - \Phi(|t|)) \\ &= 2 \cdot \left(1 - \Phi\left(\frac{|\bar{x}|}{\sigma_0/\sqrt{\tilde{k}}}\right)\right). \end{aligned} \tag{2}$$

Since we are usually not concentrating on a specific protein, but are searching for proteins with significant regulation, we apply the test to all proteins, i. e. we apply the test n times. Therefore, we should take a correction for multiple testing into account. There are various ways to treat the problem of multiple testing, for example Bonferroni (see for instance [14]), Bonferroni–Holm [7] or false discovery rate (FDR) correction [2]. It is out of the scope of this paper to discuss the correction methods for multiple testing. Here we simply use FDR correction, so that the corrected p-value is simply the one obtained from Eq. (2) multiplied by the number of applied tests, i. e. by the number of proteins measured in the experiment.

$$p_i^{(\mathrm{FDR})} = \min\{n \cdot p_i, 1\} = \min\left\{2 \cdot n \cdot \left(1 - \Phi\left(\frac{|\bar{x}|}{\sigma_0/\sqrt{\tilde{k}}}\right)\right), 1\right\}. \tag{3}$$

5.2. p-values based on a minimum number of significant regulations

The simple Z-test proposed in the previous subsection is sensitive to outliers and would yield a small p-value if a protein shows very high regulation in one replicate, even though there is more or less no regulation in the other replicates. Therefore, we propose also a stricter test that requires that the absolute value of the log2-regulation exceeds a given threshold c_α in at least m out of the k replicates. The threshold depends on the choice of the significance level α for the test. We first consider the test only for a single protein. Then

$$\binom{k}{m} \cdot \left(1 - \Phi\left(\frac{c_\alpha}{\sigma_0}\right)\right)^m \tag{4}$$

is the probability that the value of its log2-regulation exceeds c_α in at least m out of the k replicates given the null hypothesis is true, i. e. the log2-regulation factors follow a normal distribution with expected value 0 and variance σ_0^2. Since we are interested in up- and down regulations, we also need to consider the case that m out of the k log2-regulation factors are below the negative threshold $-c_\alpha$ which has the same probability (4). Therefore we should choose the threshold c_α such that

$$\alpha = 2 \cdot \binom{k}{m} \cdot \left(1 - \Phi\left(\frac{c_\alpha}{\sigma_0}\right)\right)^m \tag{5}$$

holds.

When we apply this test to all proteins, we need to take a correction for multiple testing into account. As already in the case for the simplified Z-test, we use FDR correction. Therefore, the threshold c_α should be chosen such that

$$\alpha = 2 \cdot n_m \cdot \binom{k}{m} \cdot \left(1 - \Phi\left(\frac{c_\alpha}{\sigma_0}\right)\right)^m \tag{6}$$

holds. n_m is the number of proteins which have been measured in at least m replicates. We use this value for the FDR correction, since we cannot apply the test to proteins that have been measured in less than m replicates.

From Eq. (6) we obtain the value for the threshold

$$c_\alpha = \sigma_0 \cdot \Phi^{-1}\left(1 - \left(\frac{\alpha}{2 \cdot n_m \cdot \binom{k}{m}}\right)^{\frac{1}{m}}\right). \tag{7}$$

6. EMPIRICAL EVALUATION

We have applied the above described analysis successfully in various experiments. A detailed introduction to the biological background of these experiments and the resulting data would be out of the scope this paper. Therefore, we provide results from simulated data. We assume to have $n = 5$ replicates and to measure $k = 1000$ proteins. For each protein we generate a log2-regulation factor randomly from a standard normal distribution. Then this log2-regulation factor is used as the expected value for the normal distribution for the 5 replicate measurements of the protein. The standard deviation of the normal distribution for the 5 replicates is chosen as $\sigma_0 = 0.4$. 10 % of the log2-regulation factors were deleted randomly from the table and marked as missing values.

		Threshold for significant regulation in at least k replicates		No. of significantly regulated proteins		
	$\hat{\sigma}_0$	$k = 5$	$k = 4$	$k = 5$	$k = 4$	Z-test
MAD	0.4139	1.7775	1.9419	17	29	411
S	0.4053	1.7407	1.9017	18	31	421

Tab. 4. Results for simulated data.

Table 4 shows the results for the simulated data. The estimations for the standard deviation based on MAD_5 and S_5 are close to the true value $\sigma_0 = 0.4$. Since we strictly follow the ideal model of normal distributions with the same standard deviation, S_5 yields a slightly better and smaller estimate of σ_0, so that also for the hypothesis tests more proteins will be considered as significantly regulated. When we assume an α-error of 5 % and apply FDR correction for multiple testing, the simple Z-test will identify a much larger number of proteins as significantly regulated proteins than the much stricter tests in which all 5 or at least 4 replicates must show a significant regulation. For the Z-test, a single outlier in one replicate will lead to a large absolute mean value of the log2-regulation factors of the corresponding protein and will suffice to yield a significant p-value for the Z-test. This does not apply to the tests in which the regulation of a protein must exceed a threshold in at least a certain number of replicates.

7. CONCLUSIONS AND FUTURE WORK

We have proposed a statistical approach to identify proteins that show significant regulation in mass spectrometry experiments. Our approach is not restricted to proteomics experiments, since the main characteristics – a small number of replicates and a large number of "molecular items" of interest – are typical for high-throughput experiments that yield regulation factors. An important difference between data from proteomics experiments and microarray experiments is that one often does not have a specific list of target proteins to be measured, whereas for microarrays there is fixed set of mRNA to be measured on a chip. Therefore, there are usually more missing values in proteomics than in microarray experiments. The method proposed in Section 5.2 is suitable to cope with such larger numbers of misssing values.

Results obtained from our approach with real data will be published in biological journals. Our future work will also focus on applying our approach in the context of metabolomics and genomics data.

8. SOFTWARE

The methods described in Section 5 were implemented in R, a free software environment for statistical computing and graphics [11] (see `http://www.r-project.org/`). The simple R code can be downloaded at

`http://public.ostfalia.de/~klawonn/pvalues_kybernetika.R`

ACKNOWLEDGEMENTS

This study was co-financed by the European Union (European Regional Development Fund) under the *Regional Competitiveness and Employment* objective and within the framework of the Bi^{2}SON Project *Einsatz von Informations- und Kommunikationstechnologien zur Optimierung der biomedizinischen Forschung in Südost–Niedersachsen.*

(Received August 9, 2011)

REFERENCES

[1] S. Anders and W. Huber: Differential expression analysis for sequence count data. Genome Biology *11* (2010), R106.

[2] Y. Benjamini and Y. Hochberg: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B (Methodological) *57* (1995), 289–300.

[3] D. P. Berrar, M. Dubitzky, and M. Granzow, eds.: A Practical Approach to Microarray Data Analysis. Springer, Dordecht 2009.

[4] F. P. Breitwieser, A. Müller, L. Dayon, T. Köcher, A. Hainard, P. Pichler, U. Schmidt-Erfurth, G. Superti-Furga, J.-C. and Sanchez, K. Mechtler, K. L. Bennett, and J. Colinge: General statistical modeling of data from protein relative expression isobaric tags. J. Proteome Res. *10* (2011), 2758–2766.

[5] C. Croux and P. J. Rousseeuw: Alternatives to the median absolute deviation. In: Computational Statistics (Y. Dodge J. and Whittaker, eds.), Physica *1*, Heidelberg 1992, pp. 411–428.

[6] R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York 2005.

[7] S. Holm: A simple sequentially rejective multiple test procedure. Scand. J. Statist. *6* (1979), 65–70.

[8] C. Hundertmark, R. Fischer, T. Reinl, S. May, F. Klawonn, and J. Jänsch: MS-specific noise model reveals the potential of iTRAQ in quantitative proteomics. Bioinformatics *25* (2009), 1004–1011.

[9] F. Klawonn, C. Hundertmark, and L. Jänsch: A maximum likelihood approach to noise estimation for intensity measurements in biology. In: Proc. Sixth IEEE International Conference on Data Mining: Workshops (S. Tsumoto, C. W. Clifton, N. Zhong, X. Wu, J. Liu, B. W. Wah, and Y.-M. Cheung, eds.), IEEE, Los Alamitos 2006, pp. 180–184.

[10] F. Klawonn, T. Wüstefeld, and L. Zender: Statistical modelling for data from experiments with short hairpin RNAs. In: Advances in Intelligent Data Analysis IX, Springer, Berlin 2010, pp. 79–90.

[11] R. Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna 2009, `http://www.R-project.org`.

[12] M. D. Robinson and A. Oshlack: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology *11* (2010), R25.

[13] P. J. Rousseuw and C. Croux: Alternatives to the median absolute deviation. J. Amer. Statist. Assoc. *88* (1993), 1273–1283.

[14] J. P. Shaffer: Multiple gypothesis testing. Ann. Rev. Psych. *46* (1995), 561–584.

[15] G. K. Smyth: LIMMA: Linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor (R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, eds.), Springer, New York 2005, pp. 397–420.

Frank Klawonn, Department of Computer Science, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel and Bioinformatics and Statistics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweig. Germany.

e-mail: f.klawonn@ostfalia.de and frank.klawonn@helmholtz-hzi.de