A Maximum Likelihood Approach to Noise Estimation for Intensity Measurements in Biology

Frank Klawonn
University of Applied Sciences BS/WF
Department of Computer Science
Salzdahlumer Str. 46/48
D-38302 Wolfenbuettel, Germany
f.klawonn@fh-wolfenbuettel.de

Claudia Hundertmark, Lothar Jänsch
Helmholtz Centre for Infection Research
Department for Cell Biology
Inhoffenstr. 7
D-31824 Braunschweig, Germany
claudia.hundertmark@helmholtz-hzi.de
lothar.jaensch@helmholtz-hzi.de

Abstract

Often, measurement of biological components generates results, that are corrupted by noise. Noise can be caused by various factors like the detectors themselves, sample properties or also the process of data processing and appears independently from the applied technology. When measuring two identical samples it can be observed that similar signal intensities may have inherent but varying levels of noise and that the ratio of noise decreases with increasing signal intensities. In this paper a statistical approach is introduced to estimate the noise inherent in the measured data. Based on this estimation, it is possible to provide information about the reliability of a measured signal and whether the difference between intensities is mainly caused by noise or by biological relevant cellular alterations.

1. Introduction

Cellular bio-molecules became accessible by different high throughput technologies and are the target of systematic analyses in order to define and reveal the molecular basis of life. From the biological perspective, it is very important to compare the measured intensities (amounts) for the same item (mRNA, proteins etc.) under different conditions and to conclude unambiguously (i) whether they differ significantly and (ii) to reveal the level of regulation. These conditions can be e.g. variable environmental settings (e.g. aerobe vs. anaerobe), or a different genetic background (wild type versus mutant). Systems biology nowadays would like to describe complex cellular processes quantitatively with the ultimate goal to establish predictive biological models. Thus, the confident detection of almost any relevant alteration certainly will play a decisive role in such projects.

On the one hand one should consider a signal to be noise if it can not be assigned exclusively to an individual item. This can be caused by the noise of amplification circuitries of the used detector but also be complicated by incomplete separation of individual items. On the other hand a statistical procedure should not simply reject data that might be corrupted by noise since the amount of biological samples is often limited and the applied analytical strategies are time and cost intensive. Therefore, bioinformatic strategies have to improve both the statistical characterisation of regulatory information and as a prerequisite the definition whether differences in the intensities are only caused by noisy measurements or actually by biological events.

In this paper, a statistical approach is introduced to estimate the noise inherent in the measured data. Based on this estimation, it is possible to provide information about the reliability of a measured intensity and whether the difference between intensities is mainly due to noise or caused by biological factors. In the context of microarray expression data, Bayesian approaches are very popular to estimate posterior probabilities of differential expressions [2, 3, 5] in order to determine whether observed differences in expressions are significant or not. The approach in this paper is based on the classical frequentist approach in statistics. Of course, it can be argued that for such a classical approach in principle an implicit prior can be computed, so that the Bayesian setting might be considered as the more general and explicit one. Noteworthy, our approach offers more flexibility in modelling the noise that has to be paid off by higher computational costs. Furthermore, we can also provide confidence intervals for the intensities and the probabilistic interpretation of our results is slightly different than in the Bayesian setting.

The paper is organised as follows: Section 2 describes

the considered problem and the model assumptions in more technical terms. A maximum likelihood estimator for the model parameters is derived in section 3. The confidence intervals and p-values for differences in intensities can then be computed based on the estimated model parameters (section 4). This new approach was applied to one example of a representative protein quantificiation by an iTRAQ-LC-MS/MS experiment. The results are summarised in section 5. Section 6 concludes the applied methods and results.

2. Problem Description

Basically intensities are measured for a fixed (mRNA, metabolites) or variable (proteins) set of components (items) from cells presenting different conditions that should be analysed comparatively. The biological question is which items have significantly differing intensities under these two conditions. Without any further knowledge or assumptions about the underlying biological and measurement process, it is almost impossible, to formulate a suitable model for the noise inherent in the data. This paper focuses on a more specific setting where for one condition at least two independent measurements of the intensities are available. The repetition of experiments under the same biological and analytical condition is accepted nowadays as a prerequisite in order to define standard variations and to validate the reliability of the measurements. Even in the case, when no repeated experiments are available, the method proposed in this paper can be applied, if additional information is available, for instance: For microarray data there might be a subset of genes that should not change their expressions in different conditions. Also, for many experiments only a small number of the items will change their intensities so that an estimation of the noise based on all items will lead to conservative, but still feasible results. Even if two or more independent measurements of the intensities under the same condition are available, some additional pre-processing or normalisation might be necessary. The pre-processing could also include a logarithmic transformation of the intensities, that is always part of state-of-the-art data-processing techniques used in microarray-aided mRNA expression analyses (see for instance [8, 4]). After the pre-processing, we assume that the structure of the data set is as follows:

$$\left(x_1^{(1)}, \ldots, x_1^{(k_1)}, \ldots, x_n^{(1)}, \ldots, x_n^{(k_n)}\right) \tag{1}$$

We have n items and each tuple

$$x_i^{(1)}, \ldots, x_i^{(k_i)} \tag{2}$$

represents k_i noisy measurements of the same unknown (pre-processed) intensity μ_i. The following approach was established based on the assumption that k_i is always ≥ 2, whereas it is not necessary that all k_i have the same value. Under certain ideal conditions k_i would be 4 for all $1 \leq i \leq n$ if an experiment was repeated four times. However, in some of the experiments the intensities for a few items could not be measured. In this case, the k_i-values will not be identical. Typical values for n range between 100 and a few thousand, whereas the k_i-values usually range only from 2 to 10.

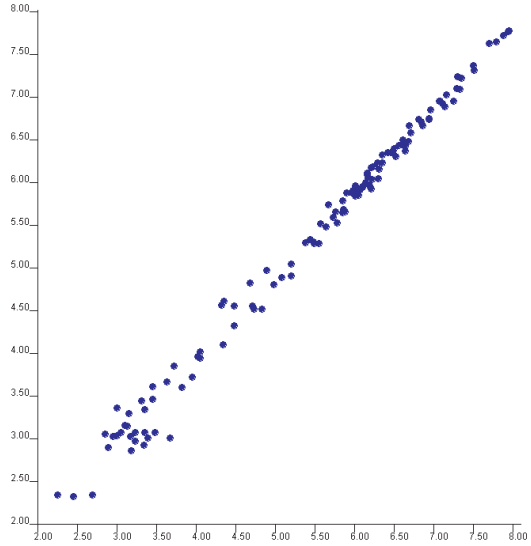

Figure 1. Normalised logarithmic intensities for a repeated experiment.

We assume that the subsample (2) originates from independent samples with normally distributed data, with unknown mean μ_i and unknown variance σ_i. Note that in most cases the values $x_i^{(j)}$ are the pre-processed intensities, so usually at least a logarithmic transformation is applied to the data. The original measured intensities cannot follow a normal distribution, since they will never yield negative values. It is, of course, impossible to make meaningful estimations for μ_i and σ_i based on a very small sample of size k_i. Furthermore, it would not really help to know the distributions for some specific intensities. From experiments we know that the variances follow a certain tendency. Small intensities are less reliable (more noisy) than larger ones. Figure 1 illustrates this effect. The (pre-processed) iTRAQ reporter intensities of 123 peptides were measured using two different reporters. In the ideal case, for each peptide the two intensities derived from the reporters would be identical. Each point in figure 1 represents the two intensities for a peptide, so that in the ideal case without noise, all points would lie on the diagonal. From this typical diagram it can be seen that the noise tends to become smaller for larger intensities. In order to take this into account, we assume that

we have

$$\sigma(\mu) \;=\; h(\mu; a, r, \lambda) \;=\; a + re^{-\lambda\mu} \qquad (3)$$

with $a, r, \lambda \geq 0$. a represents the absolute noise in the measurement. r and λ determine the intensity-dependent noise and its decrease. Our approach can be generalised easily to other noise models in the form $\sigma(\mu) = h(\mu; \theta)$ where θ is a parameter vector, in our case $\theta = (a, r, \lambda)$.

The estimation of the parameters a, r and λ based on the sample (1) requires that we estimate μ_i for each subsample (2). We only know that the values in the subsample are noisy measurements of the same intensity μ_i, but we do not know μ_i. We carry out a maximum likelihood estimation based on an expectation maximisation (EM) strategy.

3. A Maximum Likelihood Estimator Based on an Expectation Maximisation Scheme

Applying the maximum likelihood principle to estimate the unknown parameter vector $\theta = (a, r, \lambda)$ of the noise model of the previous section, leads to the likelihood

$$L\left(x_1^{(1)}, \ldots, x_1^{(k_1)}, \ldots, x_n^{(1)}, \ldots, x_n^{(k_n)} | \theta\right) \;=$$

$$\prod_{i=1}^{n} \prod_{j=1}^{k_i} \frac{1}{h(\mu;\theta)\sqrt{2\pi}} \exp\left(-\frac{(x_i^{(j)} - \mu_i)^2}{2h^2(\mu;\theta)}\right). \qquad (4)$$

The factors are simply the densities of normal distributions with mean μ_i and deviation $\sigma_i = h(\mu;\theta) = a + re^{-\lambda\mu_i}$.

The maximisation of L does not only involve the determination of the parameters a, r and λ, but also the estimation of the μ_i. Assuming the parameters a, r and λ to be fixed at the moment, we estimate the μ_i-values by maximizing the corresponding log-likelihoods. When the parameters a, r and λ are fixed, the μ_i-values can be optimised independently. Therefore, the resulting log-likelihoods are

$$\tilde{L}_i \;=\; \sum_{j=1}^{k_i} \left(-\ln(\sqrt{2\pi}) - \ln(h(\mu_i;\theta)) - \frac{(x_i^{(j)} - \mu_i)^2}{2h^2(\mu_i;\theta)}\right). \qquad (5)$$

In order to maximise $\tilde{L}_i$ it is necessary that

$$\frac{d\tilde{L}_i}{d\mu_i} \;=\; \sum_{j=1}^{k_i} \left(-\frac{h'(\mu_i;\theta)}{h(\mu_i;\theta)}\right.$$

$$\left. -\frac{-2(x_i^{(j)} - \mu_i)h(\mu_i;\theta) - (x_i^{(j)} - \mu_i)^2 h'(\mu_i;\theta)}{4h^4(\mu_i;\theta)}\right) \;=\; 0 \qquad (6)$$

holds. With $h'(\mu_i) = -\lambda r e^{-\lambda\mu_i}$ and multiplying (6) by $(a + re^{-\lambda\mu_i})^4 e^{\lambda\mu_i}$, we obtain

$$\sum_{j=1}^{k_i} \left(\lambda r(a + re^{-\lambda\mu_i})^3\right.$$

$$\left. +\frac{1}{4}(x_i^{(j)} - \mu_i)(2(ae^{\lambda\mu_i} + r) - (x_i^{(j)} - \mu_i)\lambda r)\right) \;=\; 0 \qquad (7)$$

Solving (7) for μ_i yields the maximum likelihood estimation for μ_i for fixed parameters a, r and λ. We apply a simple bisection strategy. As one boundary for bisection, we choose the mean value of the $x_i^{(j)}$. The second one is determined by a systematic search left and right of this value until the sign of (7) changes.

The optimisation of the parameters a, r and λ is carried out by a brute force algorithm, an evolution strategy with adaptive mutation rates (see for instance [1]). The fitness of a parameter combination $\theta = (a, r, \lambda)$ is given by (4), where the μ_i are determined by solving (7).

The maximum likelihood estimation uses bisection and evolution strategies. Since these methods are very general, our approach can easily be extended to other noise models of the general form $\sigma(\mu) = h(\mu;\theta)$. However, since we cannot provide an analytical solution for the parameter estimation, we have to pay the price of high computational costs (minutes up to a few hours for very large data sets on a standard PC).

As another example for a noise model, assume that the noise consists of two components, an absolute and a relative one. Then we would choose

$$\sigma(\mu) \;=\; h(\mu; a, r) \;=\; a + r\mu$$

instead of (3). The use of this formula in (6) instead of (3) can be handled in the same manner as (7).

4. Confidence Intervals and Error Probabilities for Differential Expressions

So far, we have provided a method to estimate the model parameters. Once we know these model parameters, we can derive confidence intervals for intensities as well as p-values for significant differences in intensities. As a first step, the estimation of confidence intervals is described in the following. The computation of p-values will be explained at the end of this section. When an intensity x (with noise) was measured, it is important to know the probable candidates for the true intensity. Without assuming a prior probability distribution on the intensities as in the Bayesian approach, it is impossible to specify a (posterior) probability distribution for the possible intensities. But it is possible to compute for any intensity, the probability – taking the noise into account – that it would produce an intensity like the measured one.

Assume, we want to find a "confidence interval" for the confidence level $(1-\alpha)$. Let X_μ denote the random variable with normal distribution $N(\mu, h(\mu;\theta))$. As the upper bound for the "confidence interval" we define the smallest intensity $\mu_{\max}$ that can generate an intensity lower than or equal to

the measured one with a probability of at most α. Similarly, for the lower bound we have to find the smallest intensity $\mu_{\min}$ that can generate an intensity greater than or equal to the measured one with a probability of at most α.

For the upper bound of the confidence interval we require

$$\alpha \;=\; P(X_{\mu_{\max}} \leq x) \;=\; P\left(Z \leq \frac{x-\mu_{\max}}{h(\mu;\theta)}\right) \qquad (8)$$

where Z is the standard normal distribution with mean 0 and variance 1. Therefore, we have to solve the equation

$$\Phi\left(\frac{x-\mu_{\max}}{a+re^{-\lambda\mu_{\max}}}\right) - \alpha \;=\; 0 \qquad (9)$$

for $\mu_{\max}$ where Φ is the cumulative distribution function of the standard normal distribution. This is again done by simple bisection. One boundary is chosen as x, the other is determined by searching in the entourage of x for a change of the sign of (9). Analogously, for the lower bound we require $\alpha \;=\; P(X_{\mu_{\min}} \geq x) \;=\; P\left(Z \geq \frac{x-\mu_{\min}}{h(\mu;\theta)}\right)$ which can be solved in the same way as (8).

Note that we have not assumed any prior distribution of the intensities, so that α is not the probability that the true (preprocessed) intensity lies in the interval $[\mu_{\min}, \mu_{\max}]$. It only means that a true intensity outside this range will produce a value like x with a probability lower than α.

For the question, whether two measured intensities of an item under different conditions can be considered as significantly different, a p-value for such a hypothesis can be computed. When two (preprocessed) measured intensities x_1 and x_2 with $x_1 < x_2$ are considered, it is important to know, whether the true intensity of x_1 is also smaller than the true intensity of x_2 or whether this is just an effect of noise. Let μ_1 and μ_2 denote the (unknown) true intensities of x_1 and x_2, respectively. Furthermore, assume that $\mu_1 \geq \mu_2$ holds, i.e. the true intensities are not in the same order as the measured intensities. This happens with probability

$$2 \cdot P(X_{\mu_2} \geq x_2) \cdot P(X_{\mu_1} \leq x_1). \qquad (10)$$

The factor 2 reflects that we do not consider the order (in time) in which x_1 and x_2 were measured. Without the factor 2 it would mean, we first measure the smaller value x_1 and then the larger value x_2. We have to find μ_1 and μ_2 with $\mu_2 \leq \mu_1$ such that (10) is maximised. It is obvious that μ_2 should be as large as possible, whereas μ_1 should be as small as possible. With the constraint $\mu_2 \leq \mu_1$, (10) will only achieve its maximum if $\mu_1 = \mu_2$ holds. This means, maximising (10) is equivalent to maximise

$$2 \cdot P(X_\mu \geq x_2) \cdot P(X_\mu \leq x_1) \qquad (11)$$

with the single parameter μ. (11) is equivalent to the objective function

$$g(\mu) \;=\; 2 \cdot \left(1 - \Phi\left(\frac{x_2-\mu}{h(\mu;\theta)}\right)\right) \cdot \Phi\left(\frac{x_1-\mu}{h(\mu;\theta)}\right). \qquad (12)$$

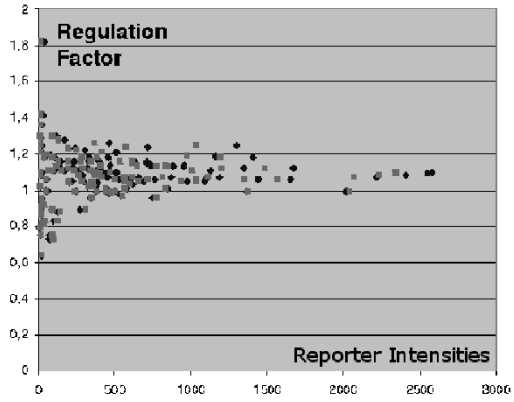

Figure 2. Comparison between regulatory information and reporter signals. Protein quantitation of two identical samples: bright dots correspond to peptides found in the sample labelled with 114.1, dark dots correspond to peptides found in the sample labelled with 116.1.

The maximisation of (12) is done again by a brute force algorithm, an evolution strategy with adaptive mutation rates. The maximum value of $g(\mu)$ is the maximum probability that two true intensities with the reverse order of x_1 and x_2 can produce values like x_1 and x_2. The result is not the probability that the true intensities for x_1 and x_2 are in the reverse order. As mentioned before, a prior distribution on the intensities is required in order to compute this probability. The computed probability is the highest possible probability that two true intensities μ_1 and μ_2 with $\mu_2 \leq \mu_1$ can result in a pair of detected intensities like x_1 and x_2.

5. Application Example

The presented noise estimation was applied to a quantitative protein analysis strategy that requires LC-MS/MS and the mentioned state-of-the-art peptide labelling strategy, known as iTRAQ [7]. Figure 2 shows the result of protein quantification of two identical samples. Consequently, the regulation factors for all peptides is expected to be nearly 1. However, the regulation factors often differ from 1 and this effect is more significant for lower reporter intensities compared to higher intensities that clearly better approximate to 1. This effect can be ascribed to intensity-dependent noise.

The iTRAQ report presented in table 1 summarises the following information: Protein ID of the identified protein P14314, the peptides (Peptide) that were sequenced

Prot. ID	Prot. Score	Peptide	Pep. Score	114.1 corr	114.1	RF	Interval	P_{err}	117.1 corr	117.1	RF	Interval	P_{err}
P14314	279,93												
		ILIEDWK	45.66	143.42	167.2	1	[92.42 ... 220.16]	0.5	337.82	368.6	2,36	[220.07 ... 516.78]	0,01
		SLEDQVEMLR	62.55	43.95	51.18	1	[26.56 ... 68.72]	0.5	58.02	65.13	1.32	[35.99 ... 90.09]	0.19
		KILIEDWK	23.98	52.67	61.42	1	[32.41 ... 81.96]	0.5	74.57	83.22	1.42	[46.99 ... 115.26]	0.14
		SLKDMEESIR	33.62	100.61	117.3	1	[64.21 ... 154.91]	0.5	133.76	147.8	1.33	[86.06 ... 205.44]	0.17
		KSLEDQVEMLR	35.59	8.61	10.14	1	[0.0 ... 15.46]	0.5	19.14	21.22	2.2	[9.09 ... 31.2]	0.5
		TVKEEAEKPER	27.1	109.62	128.7	1	[70.16 ... 168.64]	0.5	260.89	283.6	2.38	[169.6 ... 399.37]	0
		LGGSPTSLGTWGSWIGPDHDK	28.45	1	0	1	[0.0 ... 3.94]	0.5	9.14	10.13	9.14	[0.0 ... 16.24]	0.5

Table 1. Example of iTRAQ result for two different samples.

and contribute to the identification, the total (Protein Score) and the peptide associated score (Peptide Score). The information of the first four columns were obtained from the MASCOT database search engine [6] that is used for protein identification based on LC-MS/MS data routinely. The fifth and tenth column respectively contain the absolute measured iTRAQ reporter intensities for each peptide at the masses of 114.1 and 117.1 respectively, the sixth and eleventh column respectively contain the normalized and from isotopic impurities corrected intensities for each peptide at the masses of 114.1 and 117.1 respectively. The next column presents regulation factors for each peptide (RF of 114.1 set to 1 as reference). Intervals (column 8 and 13) are computed with the presented new approach: lower and upper bounds for true intensities are calculated. Within these bounds the true intensities of the measured intensities are located with the probability of a given confidence level (here: 95%). Columns P_{err} give estimated probabilities that for two measurements x_1 and x_2 with $x_1 < x_2$, the true intensity of x_1 is also smaller than the true intensity of x_2. Low probabilities (< 0.1) represent significant measurements. Based on this statistical validation our approach contributes to the biological interpretation of the data with a high confidence.

6. Conclusions

We have developed a method to improve the credibility of biological data that are acquired nowadays systematically in many laboratories. The goal of these investigations is the comprehensive and reliable detection of relevant alterations often presented as individual regulation factors for the analysed components. We designed a strategy to estimate the noise inherent in the measured data. Our approach is independent from both the type of the applied technology that provides the data and from the underlying noise model. Hence, the noise can be estimated for data from various measurement systems. Our approach can be applied to other noise models by simple variations of (3) and the computation of the corresponding model parameters as presented in section 3. We use a maximum likelihood estimator to calculate both the absolute noise in the measurement and the intensity-dependent noise, that decreases with growing intensities. After estimating absolute and intensity-dependent noise the calculated model parameters can be used to calculate error probabilities and confidence intervals for every measured intensity. Our noise estimation strategy was applied to mass spectrometry results (iTRAQ) and we are able to distinguish between significant and insignificant regulatory information and to determine an interval for the true intensity for every measured noisy intensity.

7. Acknowledgements

The authors would like to thank Tobias Reinl for critical proofreading of the manuscript and Kirsten Minkhart for excellent technical assistance.

References

[1] T. Bäck. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, Oxford, 1996.

[2] P. Baldi and A. Long. A bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.

[3] X. Cui, J. Hwang, J. Qiu, N. Blades, and G. Churchill. Improved statistical tests for differential gene expression by shrinking variance components. *Biostatistics*, 6:59–75, 2005.

[4] O. Georgieva, F. Klawonn, and E. Härtig. Fuzzy clustering of macroarray data. In B. Reusch, editor, *Computational Intelligence, Theory and Applications*, pages 83–94. Springer, Berlin, 2005.

[5] G. McLachlan, R. Bean, and L. Ben-Tovim Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22:1608–1615, 2006.

[6] P. D. C. D. Perkins, DN and J. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[7] P. Ross. Multiplexed protein quantitation in saccaromyces cerevisiae using amine-reactive isobatic tagging reagents. *Molecular & Cellular Proteomics*, 3:1154–1169, 2004.

[8] P. Spellman. Cluster analysis and display. In D. Bowtell and J. Sambrook, editors, *DNA Microarrays*, pages 569–581. Cold Spring Harbor Laboratory, Cold Spring Harbor, 2002.