A Likelihood Ratio Test for Differential Metabolic Profiles in Multiple Intensity Measurements

Frank Klawonn[1], Claudia Choi[2], Beatrice Benkert[2], Bernhard Thielen[3],
Richard Münch[2], Max Schobert[2], Dietmar Schomburg[3], and Dieter Jahn[2]

[1] Department of Computer Science
University of Applied Sciences Braunschweig /Wolfenbüttel
Salzdahlumer Str. 46/48
38302 Wolfenbüttel, Germany
[2] Institute of Microbiology
Technical University of Braunschweig
Spielmannstraße 7
38106 Braunschweig, Germany
[3] Institute of Biochemistry
University of Köln
Zülpicher Straße 47
50674 Köln, Germany

Abstract. High throughput technologies like transcriptomics using
DNA arrays or metabolomics employing a combination of gas chromatography with mass spectrometry provide valuable information about cellular processes. However, the measurements are often highly corrupted with
noise of the experimental data which makes it sometimes difficult to draw
reliable conclusions. Therefore, suitable statistical methods are needed
for the evaluation of the experimental data to distinguish changes caused
by biological phenomena from random variations due to noise. This paper
introduces a likelihood ratio test to multiple metabolome measurements.
The method was tested to differentiate differential metabolite compositions obtained from the pathogenic bacterium *Pseudomonas aeruginosa*
grown under various environmental conditions.

Keywords: Likelihood ratio test, Pseudomonas aeruginosa,
metabolomics.

1 Introduction

Various high throughput technologies enable biologists to access important
classes of cellular bio-molecules in order to gain insight into the corresponding
biological processes [1]. DNA microarray chips for measuring gene expressions
are one popular example of such high throughput technologies. Mass spectrometry is another often employed method that provides information on the presence
of molecules of interest such as proteins or metabolites [2]. The measured values
provided by these high throughput technologies are usually displayed as peak

areas or relative intensities. These values are usually compared between different conditions encountered by the analyzed organisms or tissue like variable environmental settings (e.g. aerobic versus anaerobic growth conditions). From the experimental point of view it is important to identify variations in the measured intensities under different conditions. Obtained data are usually subjected to initial normalization steps. Nevertheless, the measurements are usually corrupted by significant noise. Therefore, simply looking at the raw data and comparing intensities might lead to wrong conclusions, if the effect of the noise is ignored. Consequently, statistical methods are needed that enable the researcher to distinguish between significant differences in measured intensities and variations that are caused solely by noise. In order to estimate the influence of noise, it is necessary to have multiple measurements of the same intensities under identical experimental conditions. Unfortunately, a reliable statistical estimation requires a much larger sample size than biological experiments can deliver for reasonable costs. Repetitions of the same noisy measurement usually range between two and ten. However, by taking into account that the strength of the observed noise in the measured intensities does vary randomly, but follows a certain pattern, more reliable statistical conclusions can be drawn. Consequently, the noise estimation is not carried out separately for each analysis and measured bio-molecule, rather for all at the same time. This requires a global noise model.

Although the approach presented in this paper might be applicable to other high throughput technologies, this communication will focus on metabolome data measured by gas chromatography/mass spectrometry (GC/MS, [3]). The analysis of complete metabolomes, usually called metabolomics, has gained much attention during the past few years as an integral part of modern systems biology [4] [5]. Systems biology attempts to deduce computational models of complex cellular processes from high throughput data such as metabolomes, transcriptomes, proteomes and other related data [1].

2   Biological Context

Pseudomonas aeruginosa is a versatile soil bacterium and an important opportunistic pathogen causing persistent infection of immunocompromised patients. Particularly the lung of cystic fibrosis patients is commonly infected by *P. aeruginosa* growing as biofilm-like microcolonies. Due to this microcolony formation along with other molecular strategies *P. aeruginosa* is resistant to antibiotics, impeding patient treatment. In order to understand antibiotics resistance and the adaption to involved environmental settings, two different clinical isolates PA14 and PAO1 were grown under aerobic conditions as biofilms. Strain PA14 causes disease in a broader host range than strain PAO1 due to additional gene products [6]. Additionally planktonic cultures of strain PAO1 were collected in the early stationary phase, late stationary phase, and exponential growth phase as summarized in Table 1 (a).

5 replicates of 150 mg wet weight of the various *P. aeruginosa* strains grown under indicated conditions were collected (Table 1). Cells were harvested,

Table 1. (a) Description of all conditions, in which *P. aeruginosa* was grown and tested in this paper. (b) Number of metabolites in *P. aeruginosa*, which were different (p-value $< 10^{-10}$) between the given conditions.

Condition	Strain Growth conditions
1	PA14 aerobic, biofilm
2a	PAO1 aerobic, biofilm
2b	PAO1 anaerobic, planktonic, early stationary phase
2c	PAO1 anaerobic, planktonic, late stationary phase
2d	PAO1 anaerobic, planktonic, exponential growth phase

(a)

	1	2a	2b	2c	2d
1		17	11	26	17
2a			17	7	27
2b				24	20
2c					35

(b)

metabolites extracted, derivatized, and analyzed by GC/MS as outlined before [3]. Identification and peak areas were deconvoluted with the software AMDIS [3]. Ribitol served as internal reference standard. Identified key metabolites with significantly altered intensity pattern for the various *P. aeruginosa* strains grown under indicated conditions will be subjected to further investigation of the metabolic network and flux analysis.

3 The Noise Model

The biological experiments considered here provided data in the following form. A number of c conditions were considered where c varies between two and five in the experiments. Even if there were more than two conditions analyzed, the pair-wise comparison of conditions was the essential point for the statistical analysis. Clearly, it was sufficient to consider the situation where measurements from two conditions were available. For each condition, the intensity of each metabolite was measured k times. Typical values for k range between three and ten. Out of 176 detected metabolites from *P. aeruginosa* 107 were chemically identified. This number represents only a fraction of all metabolites present in *P. aerugi-nosa* due to the current limitation of the detection technology. In comparison, for the baker yeast *Saccharomyces cerivisiae* about 560 low molecular weight metabolites have been detected [7].

As mentioned above, it was not possible to deduce reliable statistical statements about the noise or the variance for intensity measurements for a single metabolite under a single condition tested, since a sample size of ten or less is too small. However, under the assumption that the variance followed a certain pattern depending on the true intensity, statistical inference has been carried out in a more reliable way. Before this assumption is described in more detail, necessary preprocessing steps have to be explained.

The first preprocessing step deleted all zero intensities. These intensities do not correspond to measurements, rather to missing values. Therefore, the number of available measurements for the various metabolites from one of the analyzed conditions may vary. If there were too many missing values for a certain metabolite from one analyzed condition – i.e. less than two measurements were left after removing the zeros – this metabolite was excluded from statistical analysis.

Instead of considering the measured intensities directly, a logarithmic transformation was carried out in advance. This preprocessing step is also commonly used in the context of gene expression data evaluation derived from microarray experiments [8,9]. The measured intensities in the experiments considered here ranged from values close to zero up to more than 600,000. A statistical analysis of such data tends to overemphasize the extremely large values and to neglect the smaller values. The logarithmic transformation reduces this effect. Another reason for applying the logarithmic transformation was that the original measured intensities (of one metabolite under one analyzed condition) do not follow a normal distribution, since they always yield positive values. Although the number of data for single metabolites under a single condition was too small to carry out statistical tests for the assumption of normal distribution, our interpretation of the data looked reasonable. After these preprocessing steps, the structure of the data set for one condition was as follows: $\left(x_1^{(1)}, \ldots, x_1^{(k_1)}, \ldots, x_n^{(1)}, \ldots, x_n^{(k_n)}\right)$. There were n metabolites and each tuple

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

represented k_i noisy measurements of the same unknown logarithmic intensity μ_i. As outlined before, the basic assumption was that the subsample (1) originated from independent samples with normally distributed data, with unknown mean μ_i and unknown variance σ_i. Figure 1 compares the mean logarithmic intensities on the x-axis with the empirical variances on the y-axis. Because of the small sample sizes, the variances still differ strongly. Nevertheless, there was a tendency that small (logarithmic) intensities were less reliable or more noisy than larger ones. This effect is in accordance with previous experiences and with data from other mass spectrometry experiments [10].

In the context of microarray expression data, Bayesian approaches are very popular to estimate posterior probabilities of differential expressions in order to determine whether observed differences in expressions are significant or not [11,12,13]. The approach outlined in this paper is based on the classical frequentistic approach in statistics applying a likelihood ratio test. For this test a suitable model for the noise was required taking the above mentioned decreasing noise in the logarithmic measurements into account. This was modelled by assuming that a sample of the form (1), representing k_i measurements of the same true logarithmic intensity μ_i, followed a normal distribution with unknown mean μ_i and unknown standard deviation $\sigma_i = \sigma(\mu_i) = \sigma(\mu_i; v)$. The deviation σ_i is a function of the true logarithmic intensity μ_i and a parameter vector v. In [10], it was proposed to choose $\sigma(\mu_i; v) = \sigma(\mu_i; a, r, \lambda) = a + re^{-\lambda \mu_i}$. This means the parameter vector $v = (a, r, \lambda)$ has three components. Component a represents the absolute noise independent of the true underlying logarithmic intensity μ_i, r reflects a relative portion of the noise depending on μ_i and λ determines how fast the relative portion of noise decreases. Of course, other functions are also possible to model the property of decreasing noise with increasing logarithmic intensities. Note that neither the true logarithmic intensities μ_i nor the values for the parameters a, r and λ were known in the considered experiments.
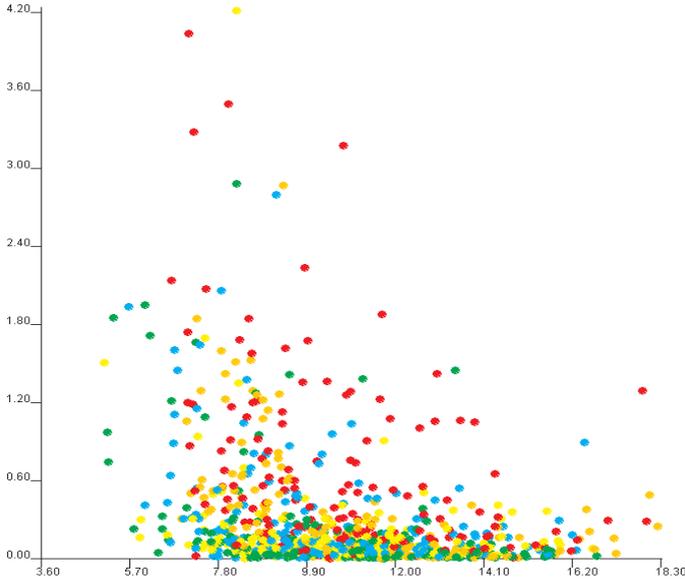
Fig. 1. Means and variances of the logarithmic intensities

4   A Statistical Test for Differential Occurrence

Considering the statistical model for the noise described in the previous section, the task was the following. There were two samples of the form (1) coming from the identified metabolite, however, measured under different conditions. The question was whether the true underlying (logarithmic) intensities μ_1 and μ_2 were identical or not. If there was enough statistical evidence that the two intensities were not identical, then this was a clear indication for a biological cause of the observed difference.

   In this section a statistical test is described that provides a p-value for the null hypothesis $\mu_1 = \mu_2$ meaning that the expression levels for the considered metabolite were identical under the two conditions versus the alternative hypothesis that true intensities differed. A low p-value meant that the null hypothesis (no differential expression) had to be rejected with high certainty. The test was derived from the application of the likelihood ratio method based on Wilks' theorem [14], stating that under certain general regularity conditions, the statistic $-2\ln T$ where

$$T = \frac{f_Y\left(Y \mid \tilde{\lambda}^{(\Lambda_0)}\right)}{f_Y\left(Y \mid \tilde{\lambda}^{(\Lambda_1)}\right)} \tag{2}$$

has an approximate χ^2-distribution. $f_Y\left(Y \mid \tilde{\lambda}^{(\Lambda_0)}\right)$ and $f_Y\left(Y \mid \tilde{\lambda}^{(\Lambda_1)}\right)$ are the likelihoods for the sample Y given the parameter vectors $\tilde{\lambda}^{(\Lambda_0)}$ and $\tilde{\lambda}^{(\Lambda_1)}$,

respectively. These parameter vectors are the maximum likelihood estimators under the constraints Λ_0 and Λ_1, respectively. In the case considered here, the parameter vector was $(a, r, \lambda, \mu_1, \mu_2)$, where μ_1 and μ_2 were the maximum likelihood estimates for the true logarithmic intensities of the metabolite from the two considered conditions. For Λ_0 the constraint $\mu_1 = \mu_2$ applies, i.e. the two true intensities are assumed to be identical. In this case the χ^2-distribution of the test statistic has one degree of freedom. The maximum likelihood estimation of the parameters involves numerical methods that are described in [10]. Here, the statistic T in (2) can be written in the form

$$T = \frac{\left(a + re^{-\lambda\mu_1}\right)^{k_1} \cdot \left(a + re^{-\lambda\mu_2}\right)^{k_2}}{\left(a + re^{-\lambda\mu}\right)^{k_1+k_2}}$$

$$\cdot \frac{\left(\prod_{i=1}^{k_1} \exp\left(\frac{(x_i^{(1)} - \mu)^2}{2\left(a + re^{-\lambda\mu}\right)^2}\right)\right) \cdot \left(\prod_{i=1}^{k_2} \exp\left(\frac{(x_i^{(2)} - \mu)^2}{2\left(a + re^{-\lambda\mu}\right)^2}\right)\right)}{\left(\prod_{i=1}^{k_1} \exp\left(\frac{(x_i^{(1)} - \mu_1)^2}{2\left(a + re^{-\lambda\mu_1}\right)^2}\right)\right) \cdot \left(\prod_{i=1}^{k_2} \exp\left(\frac{(x_i^{(2)} - \mu_2)^2}{2\left(a + re^{-\lambda\mu_2}\right)^2}\right)\right)}.$$

$(x_1^{(1)}, \ldots, x_{k_1}^{(1)})$ and $(x_1^{(2)}, \ldots, x_{k_2}^{(2)})$ are the measured (logarithmic) intensities of the considered metabolite under conditions 1 and 2, respectively. Note that the estimation of the parameters a, r and λ is based on all metabolites and not only on the metabolite under consideration, leading to a much larger sample size.

The statistical test is not only applied to one metabolite, but to a number of n metabolites in a way that several statistical tests are carried out simultaneously. A p-value considered as significant for a single metabolite could be too large in the case of multiple testing. If α is considered as a sufficiently high confidence level for a single test – i.e. p-values smaller than α lead to the rejection of the null hypothesis – a correction has to be carried out for multiple testing. Various methods are available for such corrections [15], for instance the conservative Bonferroni correction using $\frac{\alpha}{n}$ as the corrected confidence level, when n tests are carried out simultaneously. Another possible choice for the corrected confidence level is $1 - (1 - \alpha)^{1/n}$, which is applicable for two-sided hypotheses, multivariate normal statistics, and positive orthant dependent statistics, but is usually not correct in the general case [16]. In order to be on the safe side, the conservative Bonferroni correction is recommended here, especially since the likelihood ratio test is only an asymptotic, not an exact test.

5   Results and Discussion

Biofilm-like microcolonies of *P. aeruginosa* are resistant to antibiotics due to low oxygen concentration in the microcolony resulting in nearly non-growing cells amongst other factors [17]. Likewise metabolites of an aerobic biofilm PAO1 showed the highest similarity to those from anaerobic planktonic late stationary phase cells (see Table 1 (b)). Metabolomes measured for biofilms of PA14

appeared to be similar to early stationary phase cells, which indicated a more intense growth behaviour of PA14 biofilms compared to PAO1 biofilm cultures. However, the metabolome of strain PA14 in general might be different to that of strain PAO1, hence overinterpretation of these results should be avoided. With the aid of the statistical tests outlined here, we were able to find analogies between the different analyzed conditions.

Table 2. Differences in metabolite composition between late stationary phase cells versus exponential growth phase cells of *P. aeruginosa* strain PAO1 (condition 2c vs 2d) identified by the outlined likelihood ratio test. Metabolites with p-value $< 10^{-10}$ are shown. Ratio and inverse ratio show the n-fold differential occurrence of each metabolite.

Condition 2c vs 2d	Ratio r	$1/r$	p-value	Condition 2c vs 2d	Ratio r	$1/r$	p-value
Trehalose	54.38	0.02	1.8 E-14	1-6-Anhydro-beta	0.12	8.66	5.2 E-15
Quebrachitol	12.21	0.08	3.1 E-15	-D-glucose			
Mannose	8.10	0.12	7.8 E-16	Glucose-6-phosphate	0.11	8.81	3.1 E-15
Isomaltose	5.77	0.17	2.0 E-14	Hexadecanoic acid	0.11	8.92	2.6 E-15
Palmitic acid amide	0.21	4.78	4.4 E-14	Mannose-6-phosphate	0.11	9.39	3.8 E-15
1-Monostearoylglycerol	0.18	5.46	6.4 E-14	Xylose	0.09	11.46	1.2 E-15
Phenylalanine	0.18	5.49	7.1 E-15	Lyxose	0.09	11.57	<1.0 E-16
Oleic acid amide	0.16	6.34	1.3 E-15	Uridine-5'-monophos.	0.07	13.53	5.4 E-15
Xylulose-5-phosphate	0.15	6.56	7.1 E-14	Fructose-6-phosphate	0.07	14.66	4.1 E-15
Phospho-ethanolamine	0.15	6.64	8.9 E-16	Xylulose	0.07	15.30	2.1 E-11
Homoserine	0.15	6.66	5.2 E-14	Isoleucine	0.05	19.71	<1.0 E-16
Glutamic acid	0.14	7.20	3.1 E-15	Diethyleneglycol	0.05	20.49	1.3 E-15
Serine	0.14	7.22	4.4 E-16	Proline	0.05	21.12	<1.0 E-16
Threonine	0.13	7.65	<1.0 E-16	2-Monooleoylglycerol	0.04	23.76	<1.0 E-16
5-Deoxy-5-Methylthioad.	0.13	7.69	5.6 E-14	1-Monopalmitoylglyc.	0.04	26.17	<1.0 E-16
Valine	0.12	8.15	<1.0 E-16	1-Monooleoylglycerol	0.02	47.41	7.5 E-15
Ribulose-5-phosphate	0.12	8.28	1.1 E-15	Shikimic acid	0.01	121.35	1.8 E-14
Adenine	0.12	8.49	3.3 E-16	Leucine	0.01	134.68	3.8 E-15

The most pronounced difference (35 metabolites) was observed between resting cells (2c) and exponentially growing cells (2d). Details are shown in Table 2. Naturally, the activity of metabolic processes is much higher in growing cells compared to those in resting cells as reflected by in the differential metabolite profile. The di- and monosaccharides trehalose, isomaltose, and mannose were found less abundant in growing cells, since these carbon sources are usually rapidly metabolized as energy sources. Furthermore, certain sugars such as sucrose and trehalose are supposed to play a role in the adaption of *P. aeruginosa* to environmental stresses [18]. Metabolites, which are essential for amino acid biosynthesis and subsequent protein biosynthesis (leucine, proline, shikimic acid, isoleucine, valine, serine, glutamic acid, phenylalanine), for DNA and RNA formation (adenine, uridine-5'-monophosphate) and for energy production (glucose-6-phosphate, fructose-6-phosphate, mannose-6-phosphate) were found elevated in growing cells. Furthermore, lipids (palmitic acid amide, phosphoethanolamine, 2-monooleoyl-, 1-monopalmitoyl-, and 1-monooleoylglycerol) were more pronounced in growing cells, since they are essential for membrane biosynthesis.

The statistical tests for metabolome data from *P. aeruginosa* made clear differences accessible between various distinct growth conditions. Obtained results

are in good accordance to common biological interpretation. Further application of this novel statistical approach is also possible for MS/MS-peptide identification and microarray data.

References

1. Joyce, A.R., Palsson, B.Ø.: The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell. Biol. 7, 198–210 (2006)
2. Hollywood, K., Brison, D.R., Goodacre, R.: Metabolomics: current technologies and future trends. Proteomics 6, 4716–4723 (2006)
3. Strelkov, S., von Elstermann, M., Schomburg, D.: Comprehensive analysis of metabolites in Corynebacterium glutamicum by gas chromatography/mass spectrometry. Biol. Chem. 385(9), 853–861 (2004)
4. Kell, D.B.: Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov. Today 11, 1085–1092 (2006)
5. Weckwerth, W., Morgenthal, K.: Metabolomics: from pattern recognition to biological interpretation. Drug Discov. Today 10(22), 1551–1558 (2005)
6. Choi, J.Y., Sifri, C.D., Goumnerov, B.C., Rahme, L.G., Ausubel, F.M., Calderwood, S.B.: Identification of virulence genes in a pathogenic strain of Pseudomonas aeruginosa by representational difference analysis. J. Bacteriol. 184, 952–961 (2002)
7. Förster, J., Gombert, A.K., Nielsen, J.: A functional genomics approach using metabolomics and in silico pathway analysis. Biotechn. Bioeng. 79, 703–712 (2002)
8. Spellman, P.T.: Cluster analysis and display. In: Bowtell, D., Sambrook, J. (eds.) DNA Microarrays, Cold Spring Harbor Lab, pp. 569–581 (2002)
9. Georgieva, O., Klawonn, F., Härtig, E.: Fuzzy clustering of macroarray data. In: Reusch, B. (ed.) Computational Intelligence, Theory and Application, pp. 83–94. Springer, Heidelberg (2005)
10. Klawonn, F., Hundertmark, C., Jänsch, L.: A maximum likelihood approach to noise estimation for intensity measurements in biology. In: Tsumoto, S., Clifton, C., Zhong, N., Wu, X., Liu, J., Wah, B., Cheung, Y.M. (eds.) Proc. 6th IEEE Intern. Conf. on Data Mining: Workshops, pp. 180–184. IEEE, Los Alamitos (2006)
11. Baldi, P., Long, A.D.: A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. Bioinformatics 17, 509–519 (2001)
12. Cui, X., Hwang, J.T.G., Qiu, J., Blades, N.J., Churchill, G.A.: Improved statistical tests for differential gene expression by shrinking variance components. Biostatistics 6, 59–75 (2005)
13. McLachlan, G.J., Bean, R.W., Ben-Tovim Jones, L.: A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics 22, 1608–1615 (2006)
14. Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statistics 9, 60–62 (1938)
15. Miller, R.G.J.: Simultaneous Statistical Inference. Springer, New York (1991)
16. Shaffer, J.P.: Multiple hypothesis testing. Ann. Rev. Psych. 46, 561–584 (1995)
17. Borriello, G., Werner, E., Roe, F., Kim, A.M., Ehrlich, G.D., Stewart, P.S.: Oxygen limitation contributes to antibiotic tolerance of Pseudomonas aeruginosa in biofilms. Antimicrob. Agents Chemother. 48, 2659–2664 (2004)
18. Cioni, P., Bramanti, E., Strambini, G.B.: Effects of sucrose on the internal dynamics of azurin. Biophys. J. 88(6), 4213–4222 (2005)