Proteomic analysis of a lactic bacterium: level of replication and ANOVA

Olivier DAVID[1]

Christophe GITTON

Michel-Yves MISTOU

Hervé MONOD

Rapport technique 2005-3, 7 pp.

Unité Mathématiques et Informatique Appliquées
INRA
Domaine de Vilvert
78352 Jouy-en-Josas Cedex
France

[1] Olivier.David@jouy.inra.fr

Abstract

A proteomic experiment on *Lactococcus lactis* with both biological and technical replications was carried out. The data of each type of replication were analysed separately by analysis of variance. The results show that many false positives may be detected in the absence of biological replication and suggest that analysis of variance is a useful method to analyse proteomic data.

Levels of replication in 2-DE based proteomic experiments may be grouped in two types: (a) biological replication is the replication of biological samples, (b) technical replication is the replication of gels for a given biological sample. Several papers recommend to use biological replications in differential expression experiments (see e.g. [1, 2]). This recommendation is consistent with the results of several studies, which found that biological variance was larger than technical variance in proteomic experiments on plants, animals, *E. coli* (see e.g. [3, 4, 5]). The first objective of this paper is to quantify the effect of level of replication on the results of proteomic experiments on the bacterium *Lactococcus lactis*. Analysis of variance (ANOVA) has been proposed to analyse the data of differential expression experiments and has been implemented in several ways, such as gene-by-gene ANOVA or global ANOVA (see e.g. [6, 7, 8, 9, 10, 11, 12]). In a standard ANOVA, observations are assumed to follow independent normal distributions with the same variance. These assumptions may not be fully satisfied in proteomic experiments, even if a transformation of the data has been used [10]. The second objective of this paper is to study if ANOVA is a reliable method to analyse the data of proteomic experiments on *L. lactis*. The rest of this paper presents the experiment we carried out and the analyses of the data.

To compare biological and technical replications, an experiment with both types of replication was carried out. Four precultures of the same strain of *L. lactis* were used to inoculate four cultures. These cultures were carried out in four flasks, with the same medium, on the same day. One protein extract was prepared from each culture. Then, six 2-DE gels were made: Gels 1 and 2 were made from the protein extract of Culture 1, Gels 3 and 4 from the protein extract of Culture 2, Gel 5 from the protein extract of Culture 3, and Gel 6 from the protein extract of Culture 4 (Fig. 1). The experiment was performed by the same experimenter with a standard protocole. Precultures were considered as biological replications and gels made from the same protein extract were considered as technical replications. The spot volumes of 254 proteins were measured on each gel, and ranged from 106 to 237600 with a median value of 5414. 14% of volumes was missing in the initial data file, presumably because these volumes were too small to be detected. Missing observations were replaced by a small value, namely 100, which was the order of magnitude of the smallest observable volume. This experiment is similar to the uniformity trials which have been used in agriculture and in other contexts [13]. In such experiments, dummy treatments are superimposed to data obtained under uniform conditions, in order to compare design and analysis methods.

The false positive rate is the expected proportion of false positives among proteins which are not differentially expressed [14]. A false positive rate of 5% means that on average 5% of proteins which are not differentially expressed are declared positive. To assess the false positive rate for biological and technical replications, the data from both types of replication were analysed separately, with two dummy treatments applied to precultures (Fig. 1). As the experiment involved one treatment condition only, proteins declared differentially expressed were false positives.

To quantify the false positive rate for technical replications, the data of Gels 1, 2, which were made with Culture 1, and the data of Gels 3, 4, which were made with Culture 2, were analysed assuming these gels had received two treatments, denoted by 1 and 2, in the order 1, 1, 2, 2. The data were analysed using a global ANOVA (see e.g. [15]). Three factors were considered for the analysis: protein, treatment, gel. The protein factor was crossed with the other factors, and the gel factor was nested within treatment. The response y_{ptg} for protein p, treatment t, gel g was then modelled as:

$$y_{ptg} = PT_{pt} + TG_{tg} + \varepsilon_{ptg}, \tag{1}$$

where PT_{pt} was the mean response of protein p when the treatment was t and TG_{tg} was the effect of gel g within treatment t. The random errors ε_{ptg}'s were assumed to follow independent normal distributions with mean zero and variance σ^2. The response variable y was protein volume to the power λ, where λ was a parameter to estimate [16]. This transformation was used to help justify the assumptions of the model (constant variance, normality, additivity of the terms PT and TG). An alternative would have been to use a shifted logarithmic transformation [10]. The parameter λ was estimated using the function boxcox of the package MASS of R [17, 18]. The estimate of λ was equal to 0.45. Note that we did not divide the volumes of a gel by the sum of the volumes of the gel, since gels effects (TG) were naturally included in the model. To compare the relative expressions of protein p with both treatments, the parameter $\delta_p = (PT_{p2} - PT_{\bullet 2}) - (PT_{p1} - PT_{\bullet 1})$ (where $\bullet$ denotes averaging over a subscript) was considered rather than the difference $PT_{p2} - PT_{p1}$; the difference between treatment means was considered as an experimental artefact to be adjusted for. As the data were balanced, it was estimated by $\widehat{\delta}_p = (y_{p2\bullet} - y_{\bullet 2\bullet}) - (y_{p1\bullet} - y_{\bullet 1\bullet})$ (where hat denotes estimate). The p-value of $\widehat{\delta}_p$ was also calculated. It is equal to twice the probability that a random variable which follows a Student distribution with 506 degrees of freedom is larger than the absolute value of $\widehat{\delta}_p/\sqrt{\widehat{\mathrm{var}}(\widehat{\delta}_p)}$, where $\widehat{\mathrm{var}}(\widehat{\delta}_p)$ is the estimate of the variance of $\widehat{\delta}_p$. If expression differences with a p-value lower than 5% are called significant, the expected number of false positives may be large, as it may be equal to 5% times the number of proteins which are not differentially expressed. To take this multiple hypothesis testing problem into account, one possibility is to select expression differences with a q-value, rather than a p-value, smaller 5% [14]. The q-value is a measure of significance which controls the false discovery rate, i.e. the expected proportion of false positives among proteins declared positive. A false discovery rate of 5% means that on average 5% of proteins considered positive are false positives. q-values take into account the fact that many hypotheses are simultaneously tested and have been used in microarray experiments. The q-values of the $\widehat{\delta}_p$'s were calculated from the p-values using the package qvalue of R [17, 14]. As the experiment was uniform, most

q-values were expected to be large. Fig. 2 shows a histogram of the 254 q-values; 25% of proteins had a q-value smaller than 5%. So the false positive rate was poorly controlled with technical replications.

To quantify the false positive rate for biological replications, the data of one gel of Preculture 1, one gel of Preculture 2, Gel 5 and Gel 6 were analysed assuming these gels had received two treatments, denoted by 1 and 2, either in the order 1, 1, 2, 2, or in the order 1, 2, 1, 2, or in the order 1, 2, 2, 1. As there were 4 ways of choosing one gel of Preculture 1 and one gel of Preculture 2, 12 analyses were carried out in total. Each analysis was carried out with Model (1). The estimates of λ were equal to 0.25 or to 0.30. Fig. 2 shows a plot of standardized residual versus fitted value for one of the analyses. The fitted value for observation y_{ptg} is equal to $\widehat{y}_{ptg} = \widehat{PT}_{pt} + \widehat{TG}_{tg}$, and the standardized residual is equal to $(y_{ptg} - \widehat{y}_{ptg})/\widehat{\sigma}$. Thanks to the data transformation, there was not a strong relation between variance and mean. The line of points which appears in the bottom left corner of the plot is due to the fact that many spot volumes were equal to 100. Fig. 2 shows a normal quantile plot of standardized residuals from the same analysis; the residual distribution had heavier tails than a normal distribution. Fig. 2 shows a histogram of the 254 q-values from this analysis; most proteins had a q-value close to 1. For the twelve analyses carried out, the percentage of proteins with a q-value smaller than 5% ranged from 0% to 2%, with a median value of 0.4%. So except the case when the false positive rate was equal to 2%, the false positive rate was well controlled with biological replications.

Level of replication had a large effect on the results of our experiment. Thus, we recommend to choose levels of replication carefully in proteomic experiments on bacteria. In our study, ANOVA usually detected few false positives with biological replications, and was powerful enough to declare significant a number of differences of expression between precultures with technical replications. So this statistical method, although perfectible, seems to be useful for proteomic experiments.

References

[1] G. A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Supplement*, 32:490–495, 2002.

[2] M. Kathleen Kerr. Design considerations for efficient and effective microarray studies. *Biometrics*, 59(4):822–828, 2003.

[3] Victor S. Asirvatham, Bonnie S. Watson, and Lloyd W. Sumner. Analytical and biological variances associated with proteomic studies of *medicago truncatula* by two-dimensional polyacrylamide gel electrophoresis. *PROTEOMICS*, 2(8):960–968, 2002.

[4] Inmaculada Jorge, Rafael M. Navarro, Christof Lenz, David Ariza, Carlos Porras, and Jesús Jorrín. The Holm Oak leaf proteome: Analytical and biological

variability in the protein expression level assessed by 2-DE and protein identification tandem mass spectrometry de novo sequencing and sequence similarity searching. *PROTEOMICS*, 5(1):222–234, 2005.

[5] Mark P. Molloy, Erin E. Brzezinski, Junqi Hang, Michael T. McDowell, and Ruth A. VanBogelen. Overcoming technical variation and biological variation in quantitative proteomics. *PROTEOMICS*, 3(10):1912–1919, 2003.

[6] Haike Antelmann, Ron Sapolsky, Brian Miller, Eugenio Ferrari, Gopal Chotani, Walter Weyler, Alfred Gaertner, and Michael Hecker. Quantitative proteome profiling during the fermentation process of pleiotropic *bacillus subtilis* mutants. *PROTEOMICS*, 4(8):2408–2424, 2004.

[7] Nasser Bahrman, Jacques Le Gouis, Luc Negroni, Laurence Amilhat, Philippe Leroy, Anne-Lyse Lainé, and Odile Jaminon. Differential protein expression assessed by two-dimensional gel electrophoresis for two wheat varieties grown at four nitrogen levels. *PROTEOMICS*, 4(3):709–719, 2004.

[8] Paul Delmar, Stephane Robin, Diana Tronik-Le Roux, and Jean-Jacques Daudin. Mixture model on the variance for the differential analysis of gene expression. *Journal of the Royal Statistical Society, Series C*, 54(1):31–50, 2005.

[9] Andrew W. Dowsey, Michael J. Dunn, and Guang-Zhong Yang. The role of bioinformatics in two-dimensional gel electrophoresis. *PROTEOMICS*, 3(8):1567–1596, 2003.

[10] John S. Gustafsson, Robert Ceasar, Chris A. Glasbey, Anders Blomberg, and Mats Rudemo. Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data. *PROTEOMICS*, 4(12):3791–3799, 2004.

[11] M. Kathleen Kerr. Linear models for microarray data analysis: Hidden similarities and differences. *Journal of Computational Biology*, 10(6):891–901, 2003.

[12] Russell D. Wolfinger, Greg Gibson, Elizabeth D. Wolfinger, Lee Bennett, Hisham Hamadeh, Pierre Bushel, Cynthia Afshari, and Richard S. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–637, 2001.

[13] R.A. Kempton and C.W. Howes. The use of neighbouring plot values in the analysis of variety trials. *Appl. Statist.*, 30(1):59–70, 1981.

[14] J. D. Storey and R. Tibshirani. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 2003.

[15] Angela Dean and Daniel Voss. *Design and analysis of experiments.* Springer Texts in Statistics. Springer-Verlag, 1999.

[16] G. E. P. Box and D. R. Cox. An analysis of transformations. *J. Roy. Statist. Soc. Ser. B*, 26:211–252, 1964.

[17] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. ISBN 3-900051-00-3.

[18] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, fourth edition, 2002.

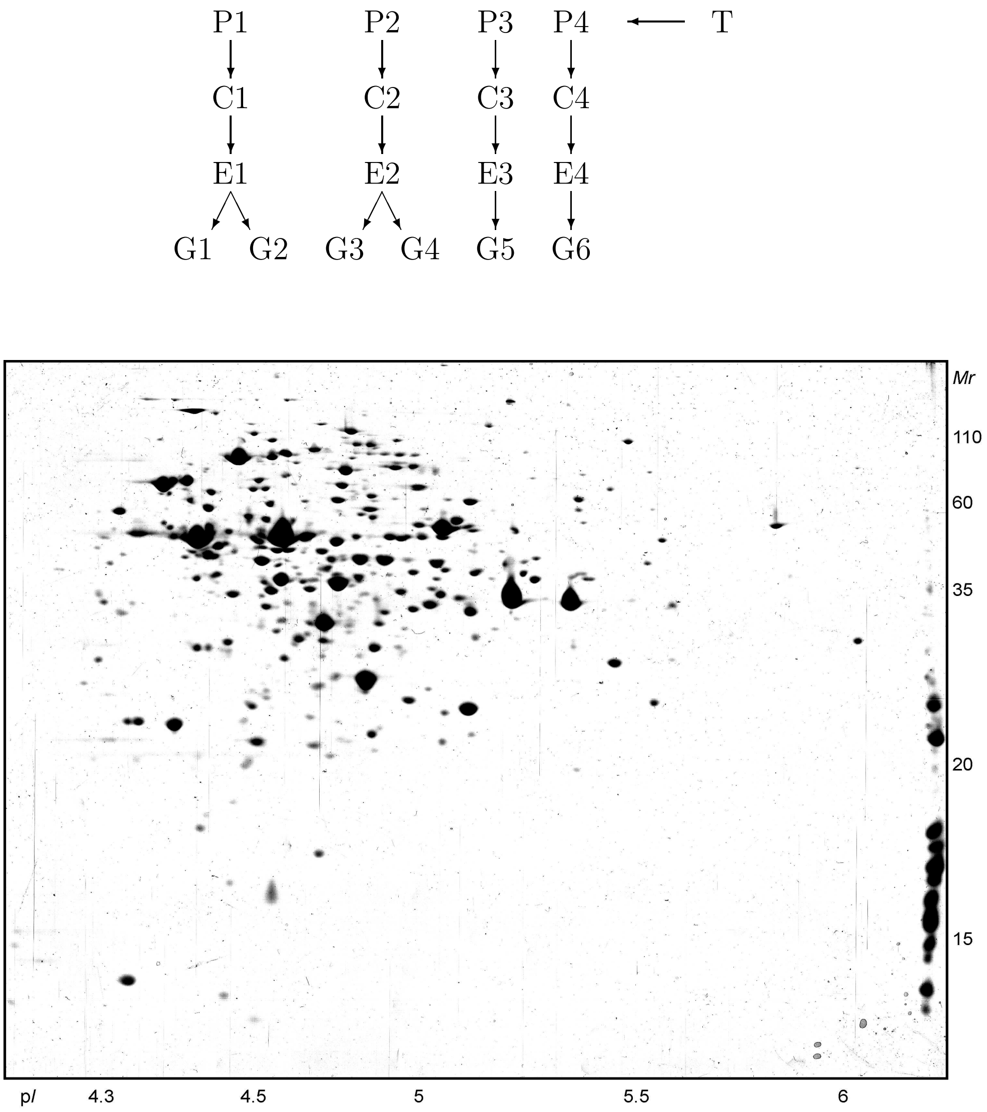

Figure 1. Schematic representation of the design of the experiment (top) and one of the six 2-DE gels of the experiment (bottom). C: culture, E: protein extract, G: gel, P: preculture, T: dummy treatment.

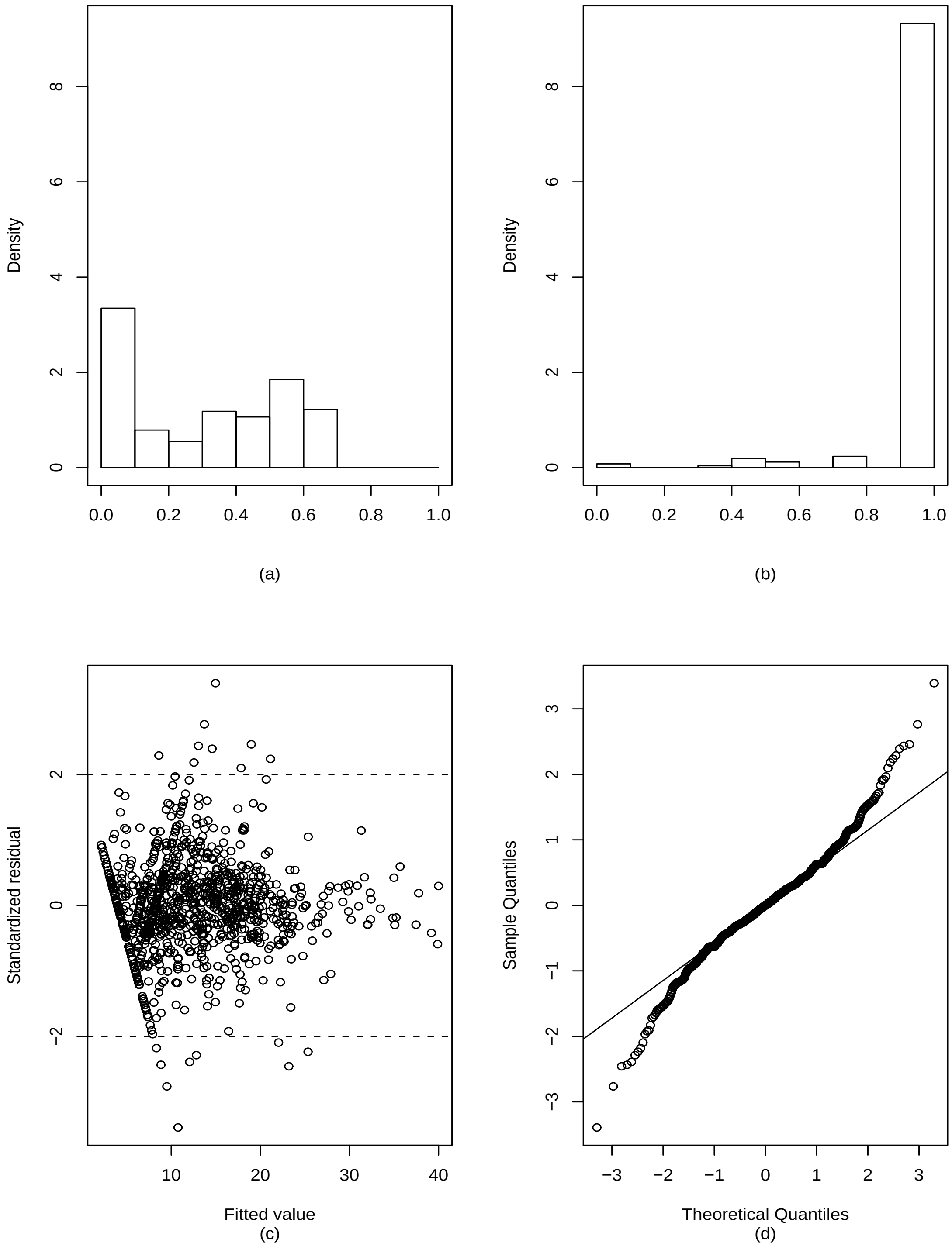

Figure 2. Density histogram of the 254 q-values for technical replications (a); density histogram of the 254 q-values (b), plot of standardized residual versus fitted value (c), normal quantile plot of standardized residuals (d), for one of the twelve analyses of biological replications.