Stockholms universitet

Statistical analysis of the effects of membrane protein overexpression in Escherichia coli

Sun Lei

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2009:2**
http://www.math.su.se

Statistical analysis of the effects of membrane protein overexpression in Escherichia coli

Sun Lei*

April 2009

Abstract

In this paper we use statistical methods to help understand how overexpression of membrane protein KDEL-receptor affects the proteome of the three host E.coli strains: BL21(DE3), C41(DE3) and C43(DE3). We analyse 411 proteins by two-way ANOVA and identify 186 proteins with a strain or treatment effect, when controlling the false discovery rate at 5%. Pairwise comparisons of the three strains at base level and stressed state respectively show that for C41 and C43, but not BL21, the majority of proteins are produced in the same amounts at both states. Chi-square Q-Q plots show that we should not assume a common error variance for the different proteins.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: lesu4833@student.su.se . Supervisor: Jan-Olov Persson.

Statistical analysis of the effects of membrane protein overexpression in *Escherichia coli*.

Sun Lei

Mars 2009

Abstract

In this paper we use statistical methods to help understand how overexpression of membrane protein KDEL-receptor affects the proteome of the three host E.coli strains: BL21(DE3), C41(DE3) and C43(DE3). We analyse 411 proteins by two-way ANOVA and identify 186 proteins with a strain or treatment effect, when controlling the false discovery rate at 5%. Pairwise comparisons of the three strains at base level and stressed state respectively show that for C41 and C43, but not BL21, the majority of proteins are produced in the same amounts at both states. Chi-square Q-Q plots show that we should not assume a common error variance for the different proteins.

Contents

Acknowledgments

I would like to express my deep and sincere gratitude to my supervisor, Research Engineer Jan-Olov Persson at Stockholm University, Department of Mathematic Statistics for his guidance and support. His wide knowledge and abstract way of thinking together with his encouraging and understanding have been of greatest value for me in order to complete the present thesis.

I am deeply grateful to Graduate student Mirjam Klepsch at Stockholm University, Department of Biochemistry and Biophysics, for her detailed and constructive comments, and for her important support throughout this work. Her presence has been essential for me.

I owe my loving thanks to my family, especially my boyfriend Tian Ren. Without his encouragement and understanding it would have been impossible for me to finish this thesis.

Stockholm, Sweden, 2009-03-11

Sun Lei

Chapter 1

Introduction

Two researchers, Samuel Wagner and Mirjam Klepsch at the Institution for Biochemistry and Biophysics at Stockholm University are analyzing the effects of KDEL-receptor overexpression in bacterium *Escherichia coli*. They are trying to understand how production of KDEL-receptor affects the cellular proteome of the host *Escherichia coli*. Proteome is the sum of all existing proteins in a cell at a given point of time.

In their experiment, they analyzed the stressed state (with KDEL-receptor overexpression) and the base level (without KDEL-receptor overexpression) for three different *Escherichia coli* strains: BL21 (DE3) pLysS, C41 (DE3), and C43 (DE3).

In this paper, we will help them to identify proteins that differ between the two different states as well as the three different strains. The analysis will be performed with two-way ANOVA separately on each of the 411 proteins. We perform multiple tests, so it is necessary to make control for false significances. They asked us to use false discovery rate (FDR) at level 0.05. FDR is the expected proportion of incorrectly rejected null hypotheses in a list of rejected hypotheses. Furthermore, we will investigate whether a same error-variance for different proteins is suitable to use in the analysis.

Chapter 2

Background and Objectives

2.1 Background

Escherichia coli (*E.coli*) is a bacterium that is commonly found in the lower intestine of warm-blooded animals. *E.coli* are not always confined to the intestine, and their ability to survive for brief periods outside the body together with the ability of growing easily and its comparatively simple and asily-manipulated genetics makes them widely used as indicator organism and the preferred choice for the high-level expression system.

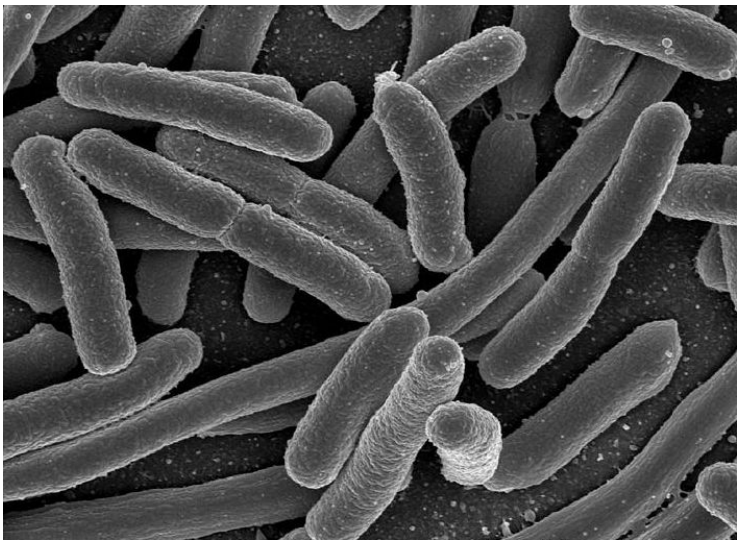

Figure 2.1: *Escherichia coli*

However, membrane protein overexpression is often toxic to cells as known, but the reasons are so far not well understand. So Wagner and Klepsch did this experiment in order to try to understand what happened to *E.coli* when overexpressing a membrane

protein - KDEL-receptor. Besides this they also wanted to understand the difference between the three *E.coli* strains BL21 (DE3), C41 (DE3) and C43 (DE3) (in the following, the three strains will be written as BL21, C41 and C43). A strain of *E.coli* is a sub-group within the species that has unique characteristics that distinguish it from other *E.coli* strains. C41 and C43 evolve from BL21 and are somehow more resistant than BL21 to the membrane protein overexpression toxicity. In theory, they should not be too different. Wagner and Klepsch set two states for the experiment. State with production of KDEL-receptor is stressed state and the state without production of KDEL-receptor is base level. [6]

The proteome of *E.coli* cells was analyzed by a method called 2-dimensional gel electrophoresis. In the first dimension, proteins were separated according to their isoelectric point and in the second dimension by their molecular weight. Wagner and Klepsch used PDQuest 8.0 from Bio-Rad to analyse the 2D-gel. Figure 2.2 shows a 2D-gel and every single black spot corresponded to a protein. They removed the irrelevant spots close to the left and right sides as well as the top part of the gel. Spots that located near actin and tubulin were overabundant and also deleted. [2]

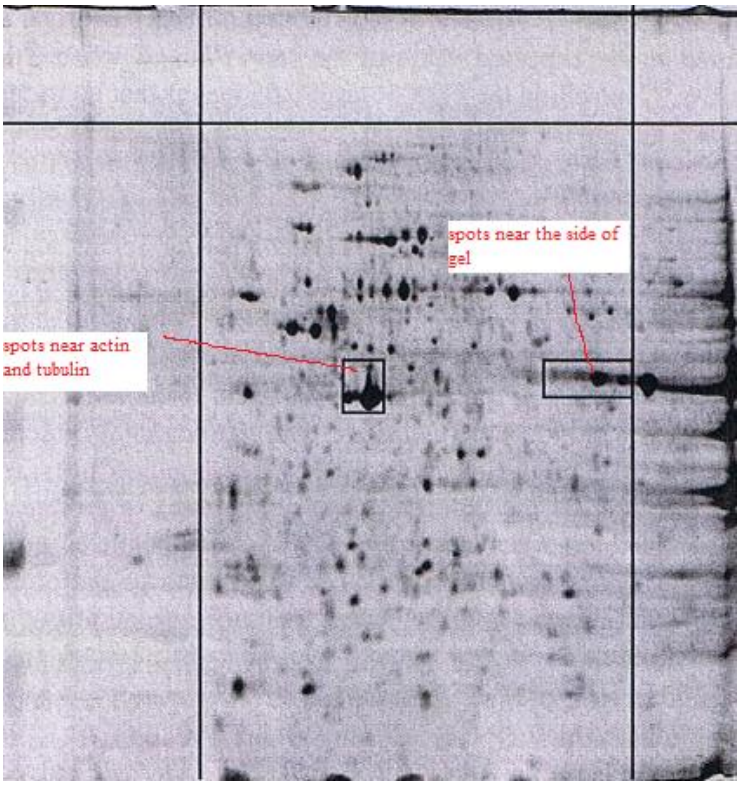

Figure 2.2: Removing irrelevant data

The software measured the optical density of each spot in the scanned gels. They did the procedure with 4 gel replicates for every strain and treatment combination and validate the spots one by one. The 522 spots in the data are all of the valid spots which they judged true. They use one normalization method for total density of valid spots according to the formula:

$$\text{normalized spot density} = \frac{(\text{raw spot density} \times \text{scaling factor})}{(\text{normalization factor})}, \tag{2.1}$$

where raw spot density is the unnormalized quantity of each spot, scaling factor is a constant and normalized factor is calculated for each gel.

2.2 Objectives

They have four objectives as Figure 2.3 shows for the experiment.

- To compare the proteomes of the three different strains at base level.
- To compare the proteomes of the three different strains at stressed state.
- To compare the base level proteome with the stressed state proteome of each strain for the production of KDEL.
- To compare the changes of the third objective between the different producing strains. [5]

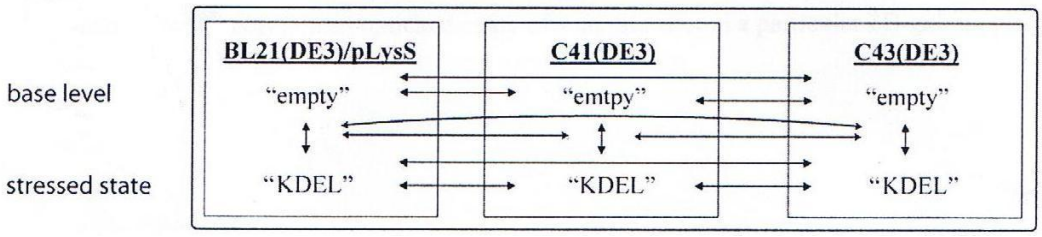

Figure 2.3: KDEL production

Besides these four objectives, we will also investigate whether there is a common experimental error for the different proteins. If so, we will use it in this analysis to strengthen the experiment.

Chapter 3

Data description

Data consists of normalized optical densities for 522 different proteins from three *E.coli* strains at two states (base level and stressed state). For each combination of strain and state, there are four replicate gels.

In theory there should be four replicates of every strain and state combination for each protein. However, none of the proteins has four replicates for every state and strain combination. There are more or less some missing values in every protein. The reason for missing value could be biological, i.e. the amount of protein is below the detective level, or technical problem or some other problem. We do not know exactly what the problems are. There are 111 proteins for which optical densities are missing in some combinations in all of the four replicate. Before doing the analysis, these 111 proteins should be eliminated since those missing values make the comparisons incomplete. After eliminating there are only 411 relevant proteins left.

In the end of paper, we will discuss a little about the eliminated data.

Chapter 4

Statistical methods and important concepts

4.1 Statistical methods

We will use two-way ANOVA model type I (only with systematic factors) to analyze the factors that affect the production level of a specific protein. Analysis of variance (ANOVA) is a collection of statistical models, in which the observed variance is partitioned into components due to different explanatory factors.

In this paper, we will study the state effects, the strain differences and the interaction between state and strain in an additive model for each protein separately,

$$log(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \sim N(0, \sigma^2), \tag{4.1}$$

where $i = 1, 2$; $j = 1, 2, 3$; $k = 1, 2, 3, 4$. With respect to the experiment we analyse:

- $log(Y_{ijk})$ is log normalised optical density (with state index i, strain index j and repetition index k);
- μ is a constant of the overall mean value;
- α_i is state parameter, $\sum_i \alpha_i = 0$;
- β_j is strain parameter, $\sum_i \beta_i = 0$;
- γ_{ij} is interaction parameter, $\sum_i \gamma_i = \sum_j \gamma_j = 0$;
- ε_{ijk} is a random disturbance and is assumed to be independent and normal distributed with expected value zero and equal variance.

The first protein is chosen as an example to illustrate ANOVA in Table 4.1. Because some values are missing in the second and fourth replicate, there are only 21 observations instead of 24. That is, there are not 4 replicates for each combination. This situation is called unbalanced. In our analysis, we use partial square sum - an attempt to compute what square sums could have been if the experiment had been balanced. Each factor is adjusted for other factors in the model, but the sum of square sums for all of the factors and residual are not necessary to be equal to the total square sum.

Table 4.1: ANOVA (analysis of variance)
Number of obs = 21 R-squared = 0.5238
Root MSE = 1.11791 Adj R-squared = 0.3651

Source	Partial SS	df	MS	F	Prob>F
Model	20.6235819	5	4.12471638	3.30	0.0330
states	0.072135557	1	0.072135557	0.06	0.8134
strains	5.46782099	2	2.7339105	2.19	0.1467
interaction	15.4385675	2	7.71928375	6.18	0.0110
Residual	18.7458364	15	1.24972242		
Total	39.3694183	20	1.96847091		

The Prob>F shows the p-values for the model 0.0330, the states 0.8134, the strains 0.1467, and the interaction 0.0110. The p-values for interaction is fairly low, indicating an interaction effect for this protein.

In order to illustrate the interaction more intuitively, we draw the following graph Figure 4.1.

As shown, the two curves are cross-cutting and not parallel. The differences between two states in different strains depart great from each other, indicating that there is interaction in the model.

The two-way ANOVA model we are using in this analysis evaluates mainly the hypothesis that there does not exist any states, strains and interaction effect, that is, $log(Y_{ijk}) = \mu + \varepsilon_{ijk}$ and p-value for the model should be larger than the significance level. If the p-value is lower than the significance level, the analysis aims to find out the differences caused by certain factors between proteins.

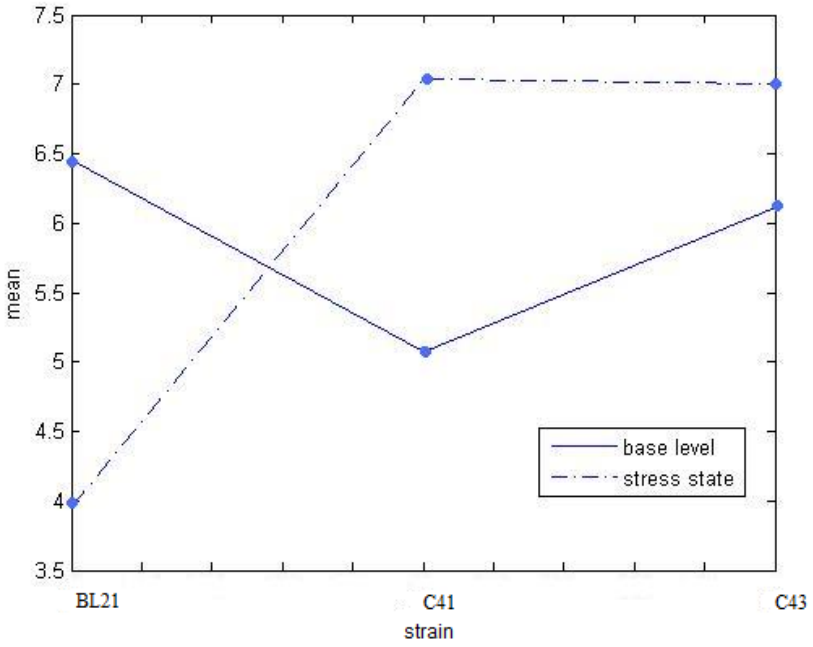

Figure 4.1: Mean values for different strains at separate states where there is interaction

4.2 Important concepts

4.2.1 Family-wise Error Rate and False Discovery Rate

In this experiment, we will perform 411 tests, so it is necessary to make control for false significances. Here we introduce two concepts for false significance controlling, Family-wise Error Rate and False Discovery Rate.

Family-wise error rate (FWER) is the probability for having at least one false positive decision among all performed tests.

Example 4.2.1. *Suppose we want to control one test at 5% significance level. If H_0 is true, it will be rejected wrongly with 5% probability.*

Example 4.2.2. *Now suppose we want to perform more independent tests, each at 5% significant level. Say 500, and all of H_0 are true. Let V denote the number of false positive decision.*

$$FWER = P(V \geq 1) = 1 - P(V = 0) = 1 - (1 - 0.05)^{500} \approx 1 \tag{4.2}$$

The probability for having at least one false positive decision among all tests is nearly hundred percent. We are nearly sure that there is at least one false positive decision in so many tests.

Now we want to control 500 independent tests at 5% FWER level instead, that is, the probability for having at least one false positive decision among all tests should be 5%. Let $\alpha^{'}$ *denote the significance level in each test and* V *is the number of false positive decision as before.*

$$FWER = P(V \geq 1) = 1 - P(V = 0) = 1 - (1 - \alpha^{'})^{500} = 0.05 \tag{4.3}$$

$$\alpha^{'} = 1 - (1 - 0.05)^{1/500} \approx \frac{0.05}{500} = 0.0001. \tag{4.4}$$

This kind of correction was developed by Italian mathematician Carlo Emilio Bonferroni. Bonferroni correction states that if an experiment is testing n independent hypotheses on a set of data at FWER level α *, then each individual hypothesis is tested at* α/n *significance level.*

The probability to reject one specific test wrongly is very low in this situation. With other words, it is hard to reject any test with such low significance level. The risk to accept wrong hypothesis is however very high. If the error from a single false rejection is not so crucial, the proportion of errors could be controlled instead. It leads us to have a look at the other measurement - FDR.

The false discovery rate (FDR), suggested by Benjamini and Hochberg (1995) is a quite new and different point of view for how the errors in multiple tests could be considered. It is the expected proportion of erroneous rejections among all rejections. In many applied problems it has been argued that the control of the FDR at some specified level is the more appropriate response to the multiplicity concern. [4]

Now considering the problem of testing simultaneously m (null) hypotheses, of which m_0 of those are true. R is the number of hypotheses rejected. Table 4.2 summarizes the situation. The specific m hypotheses are assumed to be known in advance.

The proportion of errors committed by falsely rejecting null hypotheses can be viewed through the random variable $Q = V/(V + S)$ - the proportion of the rejected null hypotheses which are erroneously rejected. Q is an unknown random variable, as V or S is unknown, even after experimentation and data analysis. When $V + S = 0$, we define $Q = 0$, as no error of false rejection can be committed. The FDR is given by the

Table 4.2: Hypotheses number

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	m_1
	$m-R$	R	m

expectation of Q,

$$FDR = E(Q) = E\{\frac{V}{V+S}\} = E(\frac{V}{R}) \tag{4.5}$$

and FWER is

$$FWER = P(V \geq 1). \tag{4.6}$$

Two important properties of FDR are easily shown:

- If all null hypotheses are true, the FDR is equivalent to the FWER.
  In this case $S = 0$ and $V = R$, so if $V = 0$ then $Q = 0$; if $V > 0$ then $Q = 1$, leading to $E(Q) = P(V \geq 1)$. Therefore control of the FDR implies control of the FWER in the weak sense. [1]

- If not all null hypotheses are true, the FDR is smaller than or equal to the FWER. We take this case into two situations,

  1. There is not any true null hypotheses,$m_0 = 0$.

     $$FWER = P(V \geq 1) = 0 \tag{4.7}$$

     and

     $$FDR = E(\frac{V}{R}) = 0. \tag{4.8}$$

  2. There are some true null hypotheses,$m_0 \neq 0$.

     Let $V = 0, 1, \ldots, m_0$ and $S = 0, 1, \ldots, m_1$,

     $$FWER = P(V \geq 1) = \sum_{i=0}^{m_0}\sum_{j=0}^{m_1} x_{ij} P(V = i, S = j), \tag{4.9}$$

     where

     $$x_{ij} = \begin{cases} 1 & \text{if } V \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

     and

     $$FDR = E(\frac{V}{V+S}) = \sum_{i=0}^{m_0}\sum_{j=0}^{m_1} \frac{i}{i+j} P(V = i, S = j). \tag{4.11}$$

Since for $i = 0$, $x_{ij} = \frac{i}{i+j}$ and for $i \geq 1$, $x_{ij} \geq \frac{i}{i+j}$,

$$\sum_{i=0}^{m_0}\sum_{j=0}^{m_1}\frac{i}{i+j}P(V=i,S=j) \leq \sum_{i=0}^{m_0}\sum_{j=0}^{m_1}x_{ij}P(V=i,S=j). \tag{4.12}$$

The result shows that $E(Q) \leq P(V \geq 1)$.

In this paper, we perform 411 tests for 411 proteins and chose FDR at level 0.05 which implies that expected 5% of tests among the rejected tests are incorrectly rejected.

The significance level used in each individual test is computed according to the Benjamini and Yekutieli procedure:

Consider testing $H_1, H_2, \ldots, H_m$ with the corresponding p-values $P_1, P_2, \ldots, P_m$. Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. Define the following multiple-testing procedure: Let k be the largest i for which $P_{(i)} \leq \frac{i}{m}q^*$; then reject all $H_{(i)}$, $i = 1, 2 \ldots, k$ and the significant level for each test should be $P_{(k)}$. [1]

Theorem 4.2.3. *For independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at q^*.*

Here we will not prove the theorem. If you are interested in the proof, please check the article "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing" by Y. Benjamini and Y. Hochberg.

4.2.2 Pooled variance

If we assume that a same phenomenon is generating random error at different situations, the different error variances can be "pooled" to express a single estimate of error variance.

We calculate the pooled variance by weighting the individual variance values with the degree of freedom of the subset at different situations. Thus, the pooled variance is given by:

$$S_p^2 = \frac{f_1V_1 + f_2V_2 + \ldots + f_iV_i}{f_1 + f_2 + \ldots + f_i}, \tag{4.13}$$

where $f_1, f_2, \ldots, f_i$ are the degrees of freedom of the data subsets at each situation, and $V_1, V_2, \ldots, V_i$ are their respective variances. [4]

An example of the pooled variance is to determine reasonable estimation of variances for each test in ANOVA model, where f_i is the degree of freedom and V_i is mean square (MS).

4.2.3 Fold change

Fold change is the ratio of the measured value for an experimental sample to the value for the control sample.

In this paper, fold change is the quota of geometric mean value of normalized optical density between stressed state and base level.

$$\text{fold change} = \frac{\overline{Y_{stress}}}{\overline{Y_{base}}}, \tag{4.14}$$

where $\overline{Y_{stress}} = (\prod_i^n Y_i)^{1/n}$ and $\overline{Y_{base}} = (\prod_j^m Y_j)^{1/m}$.

Y_i and Y_j are untransformed normalized optical densities at stressed state respective base level. In order to facilitate the expression, we take log2 fold chang and note that

$$\text{log2 (fold change)} = log2\bar{(}Y_{stress}) - log2\bar{(}Y_{base}) = \frac{1}{n}\sum_i^n log2(Y_i) - \frac{1}{m}\sum_j^m log2(Y_j). \tag{4.15}$$

Chapter 5

Analysis and Results

In this chapter, we will describe the fold changes of state effects and the comparisons of strain differences. In order to understand how the factors affect production level, we control the 411 proteins at 5% FDR level, pick out and group the significant proteins according to their different factors p-values in ANOVA models. We will then illustrate the differences between the groups. At last we will survey whether the pooled variance is suitable to use in this experiment.

5.1 Description of state effects and strain differences

In order to study the stressed state effects, we calculate the log2 fold change between stressed state and base level in each strain and the results are shown in Figure 5.1.

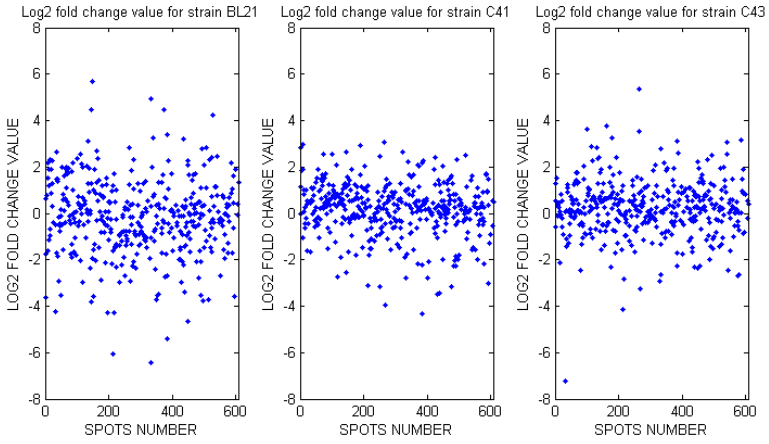

Figure 5.1: log2 Fold change for every protein between stressed state and base level in each strain

In Figure 5.1, vertical-axel scales are logarithmic. Each unit on the vertical axis corresponds to a doubling/halving. If the fold change value falls on zero, there is no effect of stressed state.

In strain BL21, the production of proteins at stress state can at most be around 64 (2^6) times as the production at base level, and lowest 0.0156 (2^{-6}) times. But the distribution of the log2 fold change values is very scattered. For some proteins, the stressed state effect is positive and for some others is negative.

In strain C41, the highest fold change value is approximately 8 and the lowest 0.0625. And majority of fold change values fall above zero, indicating that the stressed state has positive effect for majority proteins.

In strain C43, the fold change values are between 0.0078 and 32. The distribution of log2 fold change values is scattered.

From the overall, it appears that majority log2 fold change values are between -2 and 2 in all of the three graphs, indicating that for majority proteins, the productions at stressed state is between 0.25 times and 4 times as it at base level.

We then investigate the differences between strains at separate states. The pair wise comparisons are used according to the formula:

$$\text{comparison value} = \frac{\overline{Y_{strain1}}}{\overline{Y_{strain2}}}, \tag{5.1}$$

where $\overline{Y_{strain1}}$ and $\overline{Y_{strain2}}$ are geometric mean value of untransformed normalized optical densities over different strains. The expression will be transformed to log2 in Figure 5.2.

In Figure 5.2, vertical-axel scales are logarithmic. Each unit on the vertical axis corresponds to a doubling/halving. The points at zero indicate that there are not any differences between the compared strains for these corresponding proteins.

The first pair is the log2 comparison values between strain C41 and BL21 at different states. The largest comparison value is approximately 16 and the lowest 0.125 at base level. And at stressed state the highest is approximately 32 and the lowest 1/16. We see that at base level, many comparison values fall under zero, indicating that C41 is less productive than BL21 for many corresponding proteins. The distribution of values is scattered at stressed state.

The second pair is the log2 ratio between strain C43 and BL21. The highest ratio is approximately 32 and the lowest 0.125 at both states. At base level, many comparison

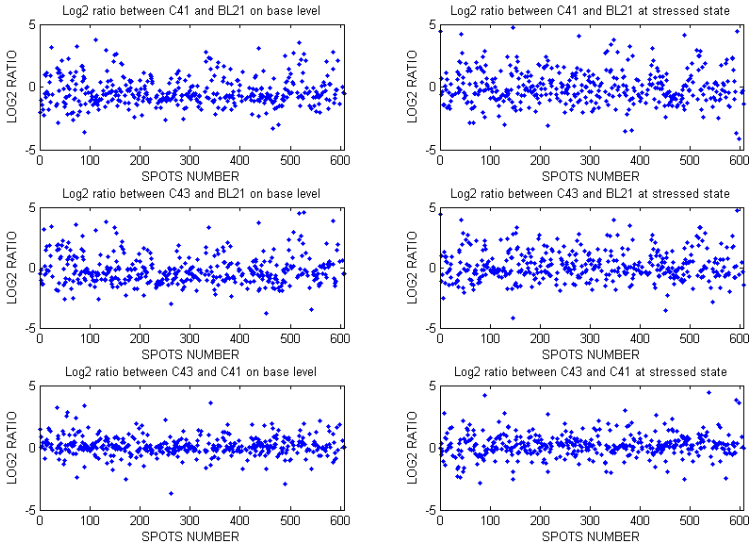

Figure 5.2: Log2 strain effect comparison at base level and stressed state for three strains

values fall under zero, indicating that C43 is less productive than BL21 for many corresponding proteins. At stressed state the distribution is scattered.

As known, C41 and C43 evolve from BL21. In theory, there should not be large difference between them. For majority proteins in the third pair, their log2 comparison values fall on zero or very near to zero. Hence we have reason to believe that for majority proteins, C41 and C43 are the same or almost the same. But at base level, the highest ratio is approximately 32 and the lowest 0.0313, and at stressed state, the highest value is approximately 32 and the lowest 0.1768.

5.2 Statistical analysis

We test 411 proteins, their corresponding model p-values in ANOVA are $P_1, P_2, \ldots, P_{411}$, with the Benjamini and Yekutieli procedure and the FDR controlled at 5% level, the significant level is therefore $P_{(k)}$ 0.0214. We have 186 significant proteins since the model p-values of these proteins are lower than 0.0214.

We use the model 4.1

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

for each protein to evaluate hypothesis that there does not exist any state effect, strains difference and interaction. So long the model p-value is lower than the significance level;

there is state effect, strain difference, interaction or any combination of the three. Now we survey the kind of effect among these significant proteins. There are no obvious ways to do that. We choose to divide the proteins in different groups according to their factors p-values.

In ANOVA we obtain the p-values for every factors of each individual significant protein. Table 5.1 summarizes the group criteria and the number of proteins is shown in Table 5.2. [3]

Table 5.1: Group criteria

	states	strains	states and strains	interaction
States p-value	<0.05	>0.05	<0.05	0-1
Strains p-value	>0.05	<0.05	<0.05	0-1
Interaction p-value	>0.05	>0.05	>0.05	<0.05

Table 5.2: Significant effects in ANOVA, number of spots

states	strains	states and strains	interaction	total
32	29	66	59	186

In Table 5.2, among 186 significant tests of proteins, 32 tests of proteins are affected (only) by states, 29 (only) by strains, 66 by both states and strains while there exists interaction in 59 proteins .

In order to illustrate the differences between different groups clearly we choose one clear out classified protein in each group. The effects/differences are shown in Figure 5.3.

In Figure 5.3, the solid line is the base level and the dotted line is the stressed state.

The first graph of Figure 5.3 shows a protein where there is almost only state effect. The two curves are far away from each other, but the differences between the strains at separate states are fairly small and the gaps between states in the three strains depart not great from each other.

The second graph shows a protein where there is strain difference. The differences between the BL21 and C41/C43 at separate states are large. The two curves are very close to each other, indicating that the state effect is not significant. Though the curves are cross-cutting and not parallel, there is no evidence for the interaction because the differences between states in separate strains are not large.

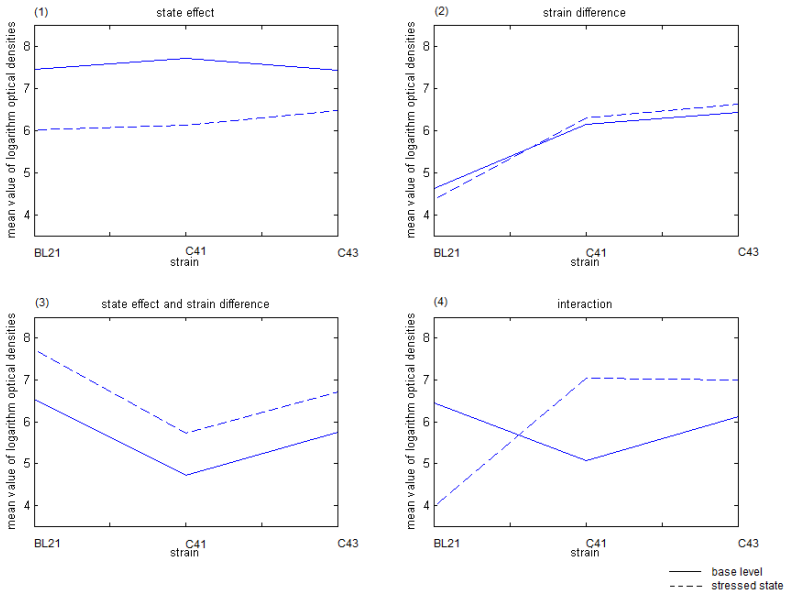

Figure 5.3: effect/differences graphic description

The third graph shows a protein in which there is state effect and strain difference. The stressed state's departure is much from base level. The differences between strains are fairly large at separate states. But the two curves are parallel, so there is no interaction between strains and states.

The fourth graph shows the interaction. The two curves are cross-cutting and not parallel, the differences between two states in different strains are great.

5.3 Analysis of pooled variance

The last but not the least interesting question is whether the pooled variance is suitable to use in this experiment. There are 411 ANOVA models, one for each protein and each model has its own variance. If there is a common variance, we can use pooled variance to estimate that, and strengthen the experiment.

The variances estimated from the 411 models have different degree of freedom (df) from 11 to 17. We group the models according to their df. If there is a common variance, for each model i with the same df f, the statistic $\frac{fS_i^2}{\sigma^2}$ should be chi-square distributed with f degree of freedom. In the statistic $\frac{fS_i^2}{\sigma^2}$, f is the degree of freedom; S_i^2 is the estimation of variance for model i; σ^2 is the common unknown-variance and is estimated as the average variance of the models with the same df f.

Since there are no more than 10 proteins for df from 11 to 13, we will only investigate the chi-square plots for the models with df from 14 to 17 in Figure5.4. Chi-square plot is a graphical tool to investigate whether the statistics $\frac{fS_i^2}{\sigma^2}$ follow chi-square distribution.

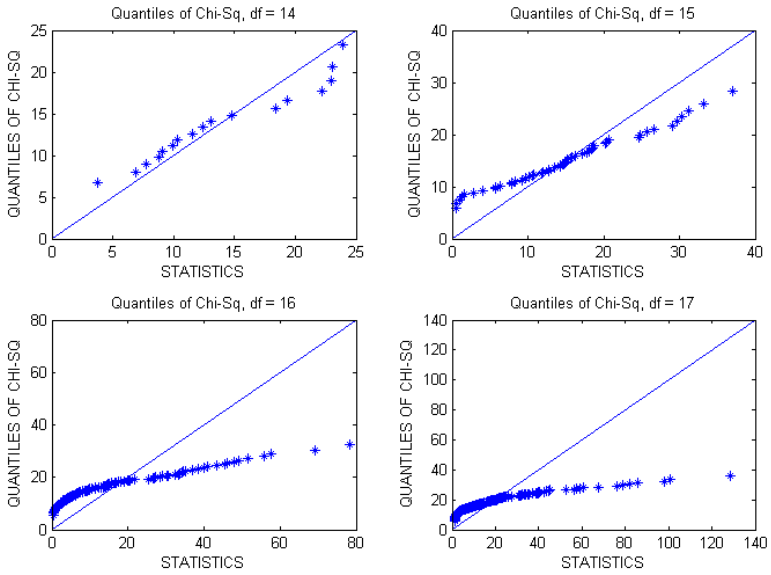

Figure 5.4: Chi-square plots for variance

X-axis are statistics $\frac{fS_i^2}{\sigma^2}$, Y-axis is quantiles of chi-square distribution. If the statistics are chi-square distributed, they should fall approximately along the 45-degree reference line. In Figure 5.4 the four graphs have a common pattern. The values are over the reference line at the beginning and then lay under the line and depart from the line, which indicates that the distributions of the models have longer tails than chi-square distribution. This situation at df 16 and 17 is particularly evident.

The models with df 16 and 17 are definitely not chi-square distributed and have not common variance, but the models with df 14 and 15 are in doubt. To confirm our judgment, we use Kolmogorov-Smirnov test to test the assumptions.

P-values for models with df 14 and 15 are 0.758 and 0.111, respectively. At 0.05 significant levels, the null hypotheses that they are chi-square distributed cannot be rejected. But whether the p-values are large enough to support the same variance cannot be promised. P-values for models with df 16 and 17 are both lower than 0.0005, with such low p-values, the models have not any common variance.

Chapter 6

Discussion

We have analyzed the different factors in the 411 proteins and illustrated the factors effect/difference. During the analysis, some things are worth to think about.

First of all there are some defects in data because there is a large number of missing values. The reason for missing value could be biological, i.e. the amount of proteins is below the detective level, or technical problem or some other problem. We eliminated 111 proteins for which values are missing in all of the four replicates. However, if the missing values are caused by the below-the-detective-level protein amount, it is very interesting and worth to study with these eliminated proteins. We could analyze which factor/factors cause such low protein amount. Unfortunately, we do not know what the reasons are for missing values, so we did the analysis in ANOVA without these 111 eliminated proteins.

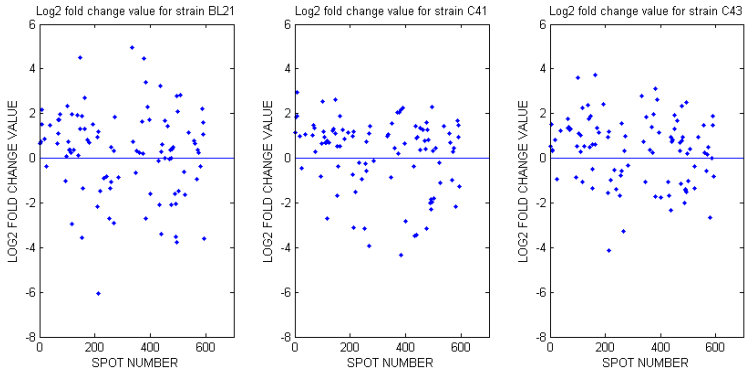

Figure 6.1: Log2 fold change values of significant proteins affected by states for each strain

When investigating the significant tests of proteins, we find an interesting phenomenon. In our opinion, the state fold change value for the significant tests affected by states should fall far away from zero. However, as Figure 6.1 shown, some points are near zero very

much. We think that the reason leading this phenomenon is the underestimation of the variances. The variances of the tests with very low fold change are probably not so small, but we underestimate them in ANOVA models, which leading the tests to be significant, and their log2 fold change values shown in the graphs are therefore close to zero.

If a common error variance existed, we would have a new significant level and the experiment would be strengthened. The phenomenon mentioned above could be improved. These tests with log2 fold change values that are very near zero would no longer be significant. Unfortunately we cannot assume a common variance for the different proteins with ANOVA. Maybe we could get a better result with other methods, but here we will not have further discussion.

Bibliography

[1] Y. Benjamini and Y. Hochberg. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.*

[2] J-F. Chich, O. David, F. Villers, B. Schaeffer, D. Lutomski, and S. Huet. *Statistics for proteomics: Experimental design and 2-DE differential analysis.* 2006.

[3] J-O. Persson. *Analysis of membrane production in E. coli-Statistical analysis.*

[4] S. Wagner. *Materials and methods.*

[5] S. Wagner and M. M. Klepsch. *Key for the analysis of membrane protein production in Escherichia coli.*

[6] S. Wagner, M. M. Klepsch, S. Schlegal, A. Appel, R. Draheim, M. Tarry, M. Hogblom, K. J. Wijk, D. J. Slotboom, J-O. Persson, and J-W. de Gier. *Tuning Escherichia coli for membrane protein overexpresion.*

Kandidatuppsats 2009:2
Matematisk statistik
April 2009

www.math.su.se/matstat

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm