Thesis for the degree of Doctor of Philosophy

Unwarping and Analysing Electrophoresis Gels

John Gustafsson

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University

Göteborg, Sweden 2005

Front cover: An illustration of a pair-wise alignment of two 2-D electrophoresis gel images. Here a gel image with a geometrically distorted spot pattern of proteins from an industrial lager yeast is unwarped into alignment with an image of the corresponding pattern of proteins from a laboratory yeast. An initial visual comparison of the protein content in the two yeast strains can be made in a pseudo-colour overlay of the two aligned gel images. (For illustration purposes, the geometrical distortion of the spot pattern in the lager yeast gel was exaggerated in advance by artificially warping the gel image based on a physiochemical model that is presented in the thesis.)

Unwarping and Analysing Electrophoresis Gels
JOHN GUSTAFSSON

ISBN 91-7291-722-9

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 2404
ISSN 0346-718X

This doctoral thesis presents results from a graduate student project that has been a collaboration with the Department of Cell and Molecular Biology at Göteborg University. The project has been financially supported by Chalmers Bioscience Programme.

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 10 00

Prepared using LaTeX
Printed and bound at the Department of Mathematical Sciences
Chalmers University of Technology and Göteborg University
Göteborg, Sweden 2005

Unwarping and Analysing Electrophoresis Gels
John Gustafsson
Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University

Abstract

Gel electrophoresis is a technology in molecular biology that is used to separate and quantify biomolecules, for example DNA and proteins. Ideally the position in a gel image corresponds to one or several properties of the molecules, such as size and net charge, but in practise the separation pattern is often warped, that is, geometrically distorted. This thesis presents methods for unwarping and analysing electrophoresis gel images. There are three main contributions in the thesis: (i) An image analysis method for simultaneous unwarping and classification of multi-track 1-D gels using a database of track profiles. The main advantage of this method is that the two analysis steps are done automatically and simultaneously. The method is motivated by and applied to pulse-field electrophoresis gels from a genotyping of potentially antibiotic-resistant bacteria. (ii) A physical model and methods for individual unwarping and pair-wise alignment of 2-D gel images. Here, the main advantages are an explicit control of image deformations and that the pair-wise alignment is automatic. The methods are evaluated on a set of protein gels from a study of the adaptation of yeast cells to saline conditions. The image alignment method is also applied and found to be robust in a comparative study of different yeast strains where the protein spot pattern differs substantially between the strains. (iii) The final main contribution is a statistical exploration of variation in quantitative 2-D gel data from three gel sets. Here, parallels to the design and analysis of microarray experiments are discussed and utilised.

Keywords

Classification, Curved field lines, Gel matching, Image alignment, Image warping, Proteomics, Pulse-field gel electrophoresis, Two-dimensional electrophoresis, Variation, Yeast.

Unwarping and Analysing Electrophoresis Gels
JOHN GUSTAFSSON
Institutionen för Matematiska vetenskaper
Enheten för Matematisk statistik
Chalmers tekniska högskola och Göteborgs universitet

Sammanfattning

Gelelektrofores är en teknologi inom molekylärbiologin som används för att separera och kvantifiera biomolekyler, t.ex. DNA och proteiner. Idealiskt motsvarar positionen i en gelbild en eller flera egenskaper hos molekylerna, såsom storlek och nettoladdning, men i praktiken är separationsmönstret oftast förvrängt geometriskt. Denna avhandling presenterar metoder för att återställa och analysera elektorforesgelbilder. Det finns tre huvudbidrag i avhandlingen: (i) En bildanalysmetod för simultant återställande och klassificering av flerspåriga 1D-geler m.h.a. en databas med spårprofiler. Den främsta fördelen med denna metod är att de två analysstegen görs automatiskt och simultant. Metoden är motiverad av och tillämpad på pulsfältselektroforesgeler från en genotypning av potentiellt antibiotikaresistenta bakterier. (ii) En fysikalisk modell och metoder för individuellt återställande och parvis bildpassning av 2D-gelbilder. Här är de främsta fördelarna en explicit kontroll över bilddeformationer och att den parvisa bildpassningen är automatisk. Metoderna är utvärderade på ett proteingelset från en studie av jästcellers anpassning till satlhaltiga förhållanden. Bildpassningsmetoden är också tillämpad och har visat sig vara robust i en jämförande studie av olika jäststammar där proteinfläcksmönstret i gelerna skiljer sig avsevärt mellan stammarna. (iii) Det avslutande huvudbidraget är en statistisk utforskning av variationen i kvantitativ 2D-geldata från tre gelset. Här diskuteras och används paralleller till designen och analysen av mikromatrisexperiment.

List of papers

This thesis is based on the work contained in the following papers:

I. Glasbey, C., Vali, L. and **Gustafsson, J.** (2005).
A statistical model for unwarping of 1-D electrophoresis gels.
Electrophoresis, 26, 4237-4242.

II. **Gustafsson, J. S.**, Blomberg, A. and Rudemo, M. (2002).
Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern.
Electrophoresis, 23, 1731-1744.

III. **Gustafsson, J. S.**, Caesar, R., Glasbey, C. A., Blomberg, A. and Rudemo, M. (2004).
Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data.
Proteomics, 4, 3791-3799.

IV. Caesar, R., Pettersson, E., Hashemi, A., **Gustafsson, J. S.** and Blomberg, A. (2005).
Comparative proteomics of industrial lager yeast reveals strain dependent differential expression of the *cerevisiae* and non-*cerevisiae* parts of their genomes.
To be resubmitted

Contents

Preface

> "You can never know everything," Lan said quietly, "and part of what you know is always wrong. Perhaps even the most important part. A portion of wisdom lies in knowing that. A portion of courage lies in going on anyway."
> *Robert Jordan (1948 -), Winter's Heart*

I start by sending my appreciation to Mats Rudemo, Anders Blomberg, and Chris Glasbey for initiating a project on image analysis of 2-D gels back in 1999. A project which later on evolved to my own graduate student project when I started my PhD studies in November 2000. This trio has also been the scientific advisors of the graduate student project, and has served as invaluable support in the diverse scientific disciplines I've encountered during the years: statistics and its mathematical framework, image analysis and molecular biology, and theoretical, applied and experimental aspects thereof.

On a more personal plane:
· Mats Rudemo, Mathematical Statistics, also my advisor during my Ph.D. studies – thanks for sharing your great experience and having a stoic patience during the many twist and turns of my journey.
· Anders Blomberg, Cell and Molecular Biology, Göteborg University – thank you for your enthusiasm and for providing data and the biologists perspective.
· Chris Glasbey, Biomathematics & Statistics Scotland – thank you for your hospitality during my two visits in Edinburgh and for sharing your intuition and approach to research.
· Aila Särkkä, co-advisor at Mathematical Statistics – warm thanks for always having time for discussions.

I would also like to thank Robert Caesar and Ellinor Peterson at the Department of Cell and Molecular Biology for guiding me in molecular biology in general and in their laboratory in particular, and for the fruitful collaboration we have had.

Looking back to my years at the Department of Mathematical Statistics, I would especially like to thank my fellow PhD students Erik Kristiansson, Anders Sjögren and Mats Kvarnström for friendship and numerous discussions over lunches.

Moreover, I would like to acknowledge the Chalmers Bioscience Programme – Bioinformatics, with among others Olle Nerman at Mathematical Statistics, for always beeing curious and equipped with a remarkable intuition.

Last but not least, I would like to send a loving thought to my family. Here a motivation is superfluous.

John Gustafsson
Göteborg
December 2005

Chapter 1

Introduction

> We must be careful not to confuse data with
> the abstractions we use to analyse them.
> *William James (1842 - 1910)*

Molecular cell biology is the study of life at the molecular level (see, for example, Alberts et al. (1998)). Here, a central dogma is that genetic information in the cell flow from DNA to RNA to proteins, see Fig. 1.1. A gene is a part of the DNA and encodes the amino acid sequence of its complementary protein. When a gene is expressed in the cell, its DNA sequence is transcribed into a RNA molecule, which acts as the blueprint for the synthesis of the encoded protein (the translation step). Proteins are the building blocks from which cells are built and they execute nearly all cell functions. For example, they catalyse reactions, take part in transmitting signals and generate movements. The study of proteins is the study of the mechanisms in the cell machinery.

Two areas of molecular biology that are encountered in this thesis are genomics and proteomics. Genomics is the scientific area that comprises mapping, sequencing and other simultaneous analyses of the whole or large parts of genome. The term proteomics was coined in the mid-1990:s and refers to the genome-wide analysis of all proteins encoded by a genome. This complete set of proteins is called the proteome with the analogy to all genes in a genome.

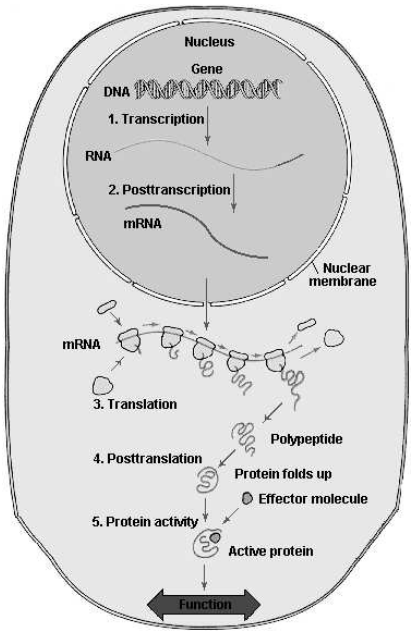


Figure 1.1: A schematic illustration of the central dogma in cell and molecular biology. ©David Ussery, DTU, Denmark

Gel electrophoresis is a technology in molecular biology that is used to separate and quantify biomolecules, for example DNA and proteins. Electrophoresis is the term for the migration of charged molecules in a solid or soluble medium caused by an applied electric field. Positively charged molecules migrate towards the cathode and molecules with negative charge move towards the anode. The velocity of the molecules depends on the force exerted by the electric field and on the frictional force between the molecules and the medium. A stronger electric field or a larger molecular charge yields a stronger force.

Three examples of gel electrophoresis techniques in genomics are pulsed-field gel electrophoresis (PFGE), amplified fragment length polymorphisms (AFLP), and denaturing gradient gel electrophoresis (DGGE). All these methods are used to characterise genetic variation and generate similar data, called multi-track 1-D gels in this thesis: each sample is represented by a vertical lane in the gel, with a density profile composed of a series of horizontal bands.

Two-dimensional (2-D) gel electrophoresis (O'Farrell, 1975) is a major technology in proteomics and is capable of simultaneously separating and quan-

tifying thousands of proteins from cell and tissue samples. The proteins are separated according to their isoelectric point in a first-dimensional separation (x axis) and their size in a second-dimensional separation (y axis). The successful separation is due to the fact that charge and weight are two independent physical properties of protein molecules. Gels with spot patterns of labelled or stained proteins are converted to digital images using a scanning device depending on the labelling/staining technique.

Two-dimensional electrophoresis has sometimes been described as an older technology. However, during the past few years we have seen some drastic refinements in proteomic technologies based on 2-D gels, where the major advances are (i) the improved 2-D pattern standardisation by the use of immobilised pH gradients in the first-dimensional separation (Görg et al., 2000) and (ii) the improved downstream identification of separated proteins by mass spectrometry (Pandy and Mann, 2000). The 2-D technology is also a powerful tool in the analysis of post-translational modifications of proteins (see, for example, Rabilloud (2002)).

The image analysis of the scanned 2-D gels typically comprises the following two main steps. An initial step with the detection and quantification of spots in each gel image, and a subsequent matching step where spots in different gels that correspond to the same protein are matched (see, for example, Pleissner et al. (2001); Dowsey et al. (2003)). Both these steps are non-trivial and often require time-consuming manual interference because of problematic spots and geometric distortions of the spot patterns that are inherent to the casting, polymerisation, and running procedure of the gels. These problems are referred to as the "image analysis bottle-neck" and have consequences for a comprehensive statistical analysis (Voss and Haberl, 2000).

The objective of a statistical analysis of quantified protein spot data can for example be to identify differentially expressed proteins or to perform a cluster or discriminant analysis of proteins or samples. Quantitative 2-D gel data typically comprise many variables (proteins) but few observations (experimental replications). This is also the case in the genome-wide measurement of mRNA levels with microarrays. There has been an extensive work in the statistical community on the design and analysis of gene expression experiments using microarrays (see, for example, Parmigiani et al. (2003); Speed (2003)). It is worth noticing that the statistical challenges in these two types of genome-wide analyses – microarrays at the mRNA level and 2-D gels at the protein level – are similar.

Mathematically, image warping is a function which deforms images; all positions in one image plane are mapped to positions in another plane. It has many applications in the field of image analysis (Glasbey and Mardia, 1998): for example to remove optical distortions caused by camera lenses or viewing perspective, to register an image with a template, or to bring two or more images into alignment. With the availability of fast desktop computers, which allow gel image matching at the pixel level (pixels are the picture elements which make up an digital image), there has been increasing interest from the 2-D gel community in methods from the image warping field (see, for example, Dowsey et al. (2003)).

The aim of this thesis is to improve the design and analysis of gel electrophoresis experiments. The emphasis of the work has been on gel image analysis, with the aim to develop image warping methods for multi-track 1-D gels and proteomic 2-D gels. The emphasis has also been on statistical analysis of proteomic 2-D gel data, with the aim to explore the variability statistically and to utilise parallels to the design and analysis of genomic microarray experiments.

<table>
<tr><th></th><th>Multi-track 1-D gels</th><th>2-D gels</th></tr>
<tr><td>Image analysis</td><td rowspan="2">Paper I,
complementary study 1</td><td>Paper II, Paper IV</td></tr>
<tr><td>Statistical analysis</td><td>Paper III,
complementary study 2</td></tr>
</table>

Figure 1.2: A schematic classification of the topics in the papers and complementary studies included. A column in the table corresponds to one type of electrophoresis gel and a row corresponds to a step in the analysis of the gel experiment.

Chapter 2

Summaries of papers included

> An expert is a person who has made all the mistakes
> that can be made in a very narrow field.
> *Niels Bohr (1885 - 1962)*

Paper I: A method for simultaneous unwarping and classification of multi-track 1-D gels

This paper presents a method that automatically and simultaneously unwarps and classifies track profiles in multi-track 1-D gels using a databases of profiles of known biological specimen.

The method is derived from, also in the paper, a formulated statistical model, which relates track profiles in 1-D gels to databases of profiles. The warp in each track is decomposed into a component that is parametrised and smooth over the whole gel, and a component that is modelled as stochastic and uncorrelated between tracks. Further, pixel-wise differences are modelled by a mixture of two normal distributions, which results in robust estimates. Maximum likelihood estimates are computed efficiently by an algorithm that alternates between dynamic time warping to align individual tracks and generalised-least-squares regression to ensure that the warp is smooth between the tracks.

Three pulse-field electrophoresis gels from a genotyping of potentially antibiotic-resistant *E. coli* O157 strains are used to illustrate the method. Here, the method also facilitates manual identification of new genotypes.

Paper II: A two-step image warping method for alignment of 2-D gels

The objective in this paper is to develop a method that aligns the spot pattern in a group of 2-D gel images. The proposed method works at the image level and uses image warping techniques. A major cause for geometric distortion of separation patterns in electrophoresis gels sometimes seems to be current leakage in the experimental procedure. A physical model of warping of spot patterns in 2-D gels caused by current leakage is therefore formulated.

The method consists of two image warping steps. In the first warping step, gels are treated individually and the physical model is used to correct for the effect of suspected current leakage. For each gel, the physical model is fitted to the gel image and then the gel image is unwarped according to the fitted model. In the second warping step, the gel images are treated as a group. A simple approach is used to align images in the group. The current leakage corrected image with least estimated current leakage is selected to be a reference image, and the other current leakage corrected images in the group are aligned to this reference in pair-wise image alignments. A statistically formulated warping method based on the one proposed by Glasbey and Mardia (2001) is used for the pair-wise image alignment. An original feature of the suggested pair-wise alignment method is that the pixel values are logarithmically transformed. This action enhances structures in the gel image background that often are reproducible, and might contribute to the observed robustness of the alignment method.

The performance the proposed method is evaluated on a set of ten gels from a study of the adaptation of yeast cells to saline conditions. For example, the matching performance of a commercially available software package is improved when given aligned images as compared to original ones.

Related work on pixel-based matching of 2-D gel images has been done by Conradsen and Pedersen (1992), Smilansky (2001) and Veeser et al. (2001). The largest conceptual difference of our method compared with these methods is that we keep explicit control of the distortions introduced

in the warping. It may be noted that we were not aware of the work by Veeser et al. (2001) at the time of the writing of the paper.

Paper III: A statistical exploration of variation in quantitative 2-D gel data

This paper is concerned with a thorough statistical characterisation of the variation in quantitative 2-D gel data. There have been several studies of the variation in and reproducibility of quantitative 2-D gel data (see, for example, Garrels (1989); Burstin et al. (1993); Blomberg et al. (1995); Voss and Haberl (2000); Mahon and Dupree (2001); Molloy et al. (2003)). This study is more focused on statistical aspects of the variation.

Quantitative 2-D gel data from three gel sets are examined and statistical techniques that are used to analyse the variation in the data are (i) variance versus mean stabilising transformations that have been used in the analysis of microarray data (Huber et al., 2002; Durbin et al., 2002), and classical techniques of analysing them (Tukey, 1957; Box and Cox, 1964); (ii) analysis of variance of sample variances (Scheffé, 1959) to examine the dependence of sample variances on proteins, strains and growth conditions in the experimental design; and (iii) exploration of externally Students residuals (see, for example, Draper and Smith (1998, p. 208)) to examine normality and to find systematic trends in the measurement errors.

We find that the variance versus mean dependence in data is stabilised by adding a small value before applying a logarithmic transformation, that there is a remaining clear variance heterogeneity for different proteins and that there is an approximate normality in transformed data. Further, we observe spatial correlations among the measurement errors.

These statistical observations both point at possible systematic errors in the preceding experimental and image analysis steps, and guide the choice of methods in a subsequent statistical analysis. These aspects are exemplified and discussed in the paper. We also try to emphasise the similarities with the analysis of microarray data.

Paper IV: An application of the pair-wise alignment method for 2-D gels

The aim of this paper is to analyse the proteome of three industrial lager yeast strains. One of the measurement technologies used in the study is 2-D gel electrophoresis and a main finding is that the protein content in all three lager yeast strain is very different from a standard laboratory strain of *S. cerevisiae*. For qualitative analysis of gel images using pseudo colour overlay, images where first warped into alignment using the pair-wise alignment method in Paper II. Even though the protein spot pattern differs substantially between the strains, the method automatically finds proper gel image alignments.

Chapter 3

Complementary studies

> The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them.
> *William Bragg (1862 - 1942)*

3.1 Two variants of the simultaneous analysis of 1-D gels

This section presents a feasibility study of two variants of the simultaneous unwarping and classification of multi-track 1-D gels. The aim of this study is to address the problem pointed out in the third paragraph of Section 4 of Paper I. There it is concluded that although each of the three steps in the suggested algorithm finds a globally optimal solution to a sub-problem, there is no guarantee that the algorithm as a whole finds the globally optimal solution to the full problem. Moreover, it is discussed that stochastic algorithms, such as simulated annealing, may reduce the risk of finding a suboptimal solution. This complementary study investigates such a stochastic algorithm.

The investigated stochastic algorithm is based on a generalisation of dynamic programming (DP) that has been suggested by Glasbey (2005) for restoration of synthetic aperture radar images. The algorithm is inspired by simulated annealing (SA). However it is not a true SA algorithm and therefore it is here referred to as a pseudo-SA algorithm. There are no theoretical results that guarantee that the pseudo-SA algorithm finds the

globally optimal solution.

Problem formulation

The image analysis task and the data are the same as in Paper I. The task is to unwarp and classify tracks in multi-track 1-D gel images using a database of track profiles. The data are three pulse-field electrophoresis gels from a genotyping of potentially antibiotic-resistant bacteria (Vali et al., 2004) and a database with five track profiles. The gel images are preprocessed as described in Paper I and we let Y_{ij} denote the value of picture element (pixel) in row i and column/track j of a preprocessed gel image. Similarly, we let μ_{ij} denote the value of element in row i and column/class j of the database image. The database and preprocessed tracks from the three gels are shown in Fig. 3.1.

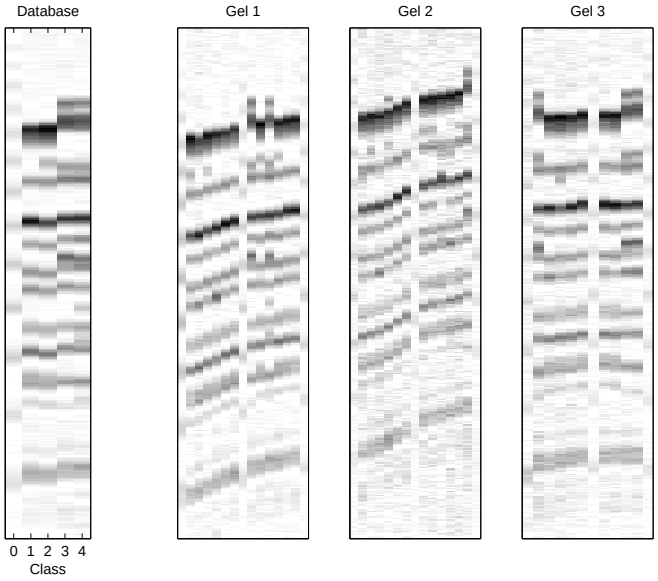


Figure 3.1: Database and preprocessed tracks from the three gels.

A mathematical formulation of the image analysis task is that we seek to find a warp f and classification l such that Y_{ij} and μ_{f_{ij},l_j} are similar in some sense for all i and j. More specifically, the warp f is an array of real numbers, $f = \{f_{ij} \in \mathbb{R}; i = 1, ..., I, j = 1, ..., J\}$ and the classification l is a vector of integers, $l = \{l_j \in \{0, ..., 4\}; j = 1, ..., J\}$. When f_{ij} is not an integer we use linear interpolation to obtain $\mu_{f_{ij}}$ and when f_{ij} lies outside the image domain $\mu_{f_{ij}}$ is set to the a pixel value corresponding to white. To measure the similarity we use the simple sum of squared differences which can be interpreted as a log-likelihood in a statistical framework where the pixel-by-pixel differences are modelled to be independently and normally distributed with mean zero and with a common variance. To control rough distortions we penalise large deviations of the first order derivatives (differences) of f from the first order derivatives of the identity warp.

The resulting optimisation problem is to find the warp f and classification l that minimise

$$S(f, l, \lambda) = \sum_{j=1}^{J} \sum_{i=1}^{I} ((Y_{ij} - \mu_{f_{ij},l_j})^2 + \lambda_1 (f_{ij} - f_{i-1,j} - 1)^2 + \lambda_2 (f_{ij} - f_{i,j-1})^2) \quad (3.1)$$

where $\lambda = (\lambda_1, \lambda_2)$ is a vector with two parameters that determines the relative weight between the similarity criterion, the first term, and the two distortion penalties, the second and third terms, respectively.

To avoid wraparound, the warp is restricted to be monotonically increasing in i, that is, down the track. Further, to speed up computations, the change in f going one pixel down the track is restricted to be less than 2 pixels in the database image. Mathematically, these two conditions can be expressed as

$$0 < f_{i+1,j} - f_{ij} < 2$$

In order to use dynamic programming we need to make f discrete. This is done by using a step size $\Delta = 0.25$. Therefore, in practise, the algorithm searches among seven possible steps $(0.25,\ 0.5,\ 0.75, ..., 1.75)$. Finally, to speed up computation even more, deviation from the identity warp is restricted to be less or equal to 100 pixels, that is, $|f_{ij} - i| \leq 100$, for all j.

Two algorithms to solve the optimisation problem

A deterministic algorithm based on dynamic programming

Fix λ.

0) (Initialisation) For track $j = 1, ..., J$, do the following.

   a) For every possible classification of the track, $l_j = 0, ..., 4$, use DP to minimise $S(f, l, 0)$ with respect to $f_{.j}$, that is, simply minimise

$$\sum_{i=1}^{I} (Y_{ij} - \mu_{f_{ij,l_j}})^2$$

   b) Set $f_{.j}$ and l_j to the values that minimised $S(f, l, 0)$.

1) For each track j in turn, do the following.

   a) For each classification of the track, l_j, use DP to minimise $S(f, l, \lambda)$ with respect to $f_{.j}$ with all other f's and l's fixed, that is, minimise

$$\sum_{i=1}^{I} ((Y_{ij} - \mu_{f_{ij,l_j}})^2 + \lambda_1 (f_{ij} - f_{i-1,j} - 1)^2 + \\ + \lambda_2 (f_{ij} - f_{i,j-1})^2 + \lambda_2 (f_{i,j+1} - f_{i,j})^2)$$

   b) Set $f_{.j}$ and l_j to the values that minimised $S(f, l, \lambda)$.

2) Repeat step (1) until convergence.

A stochastic algorithm based on dynamic programming and pseudo-simulated annealing

This algorithm is an analogue to the stochastic algorithm suggested by Glasbey (2005) for restoration of SAR images. Basically, we seek to find the warp f and classification l by minimising

$$\tilde{S}(f,l,\lambda,e) = \sum_{j=1}^{J}\sum_{i=1}^{I}((Y_{ij} - \mu_{f_{ij},l_j})^2 + e_{ij,f_{ij},l_j} + \\ +\lambda_1(f_{ij} - f_{i-1,j} - 1)^2 + \lambda_2(f_{ij} - f_{i,j-1})^2)$$

as in the deterministic algorithm, but for each iteration (sweep over the tracks), we simulate new $e_{ijfl} \sim U[0,T]$, and let the temperature T decrease slowly as the iterations proceed.

Fix λ. Set the start and final temperatures, T_S and T_F, and the number of iterations, N. Calculate the cooling constant ϕ.

0) Initialise $T = T_S$, and initialise f and l as in step (0) of the deterministic algorithm.

1) Simulate $e_{ijfl} \sim U[0,T]$ for all i, j, f and l.

   For each track j in turn, do the following.

   a) For each classification of the track, l_j, use DP to minimise $\tilde{S}(f,l,\lambda,e)$ with respect to $f_{\cdot j}$ with all other f's and l's fixed, that is, minimise

$$\sum_{i=1}^{I}((Y_{ij} - \mu_{f_{ij},l_j})^2 + e_{ij,f_{ij},l_j} + \lambda_1(f_{ij} - f_{i-1,j} - 1)^2 + \\ +\lambda_2(f_{ij} - f_{i,j-1})^2 + \lambda_2(f_{i,j+1} - f_{i,j})^2)$$

   b) Set $f_{\cdot j}$ and l_j to the values that minimised $\tilde{S}(f,l,\lambda,e)$.

2) Reduce the temperature to $T \cdot \phi$, and repeat step (1) until $T \leq T_F$.

We set the start temperature T_S such that the expected value of the random term for this temperature is the maximal squared difference possible, that is we set

$$T_S = 2(\max(Y_{ijg}, \mu_{ij}))^2$$

(Recall that the pixel values in the track profiles are preprocessed ones and not the original unsigned integers.) To set a suitable final temperature, we can establish that the smallest non-zero change in S due to a change of f in the vertical dimension, for example, is

$$\lambda_1 \Delta |f_{ij} - f_{i-1,j} - 1|$$

The smallest non-zero value of the final factor of this expression is Δ. Therefore, we set

$$T_F = 0.5 \min(\lambda_1, \lambda_2) \Delta^2$$

in order for the random term to be less than the change in the distortion penalty terms when changing f. The performance of the algorithm for different cooling schedules is investigated by setting the number of iterations be equal to $8, 16, 32, ..., 1028$.

A cross validation technique to choose weight parameters

The cross-validation described below is constructed to emphasise validation of the warping performance.

1) (Leave-one-out) For each track j', leave it out and minimise

$$\sum_{j \neq j'} \sum_{i} ((Y_{ij} - \mu_{f_{ij}, l_j})^2 + \lambda_1 (f_{ij} - f_{i-1,j} - 1)^2 + \lambda_2 (f_{ij} - f_{i,j-1})^2)$$

with respect to f and $l_1, ..., l_{j'-1}, l_{j'+1}, ..., l_J$ using the algorithm to be evaluated.

2) (Validation step) Calculate the error when predicting track j' with the most similar warped version of the database tracks when the estimated $f_{.j'}$ is used as warp. This prediction error is obtained by minimising

$$E_{j'} = \sum_{i} (Y_{ij'} - \mu_{f_{ij'}, l'_j})^2$$

with respect to $l_{j'}$. Hence, the classification of the left-out track is not based on its true class but on minimisation of the prediction error.

Repeat steps (1) and (2) for all tracks ($j' = 1, ..., J$) and, thereafter, add up to get the cross-validation error for the whole gel.

$$E = \sum_{j'=1}^{J} E_{j'}$$

Results and discussion

The results of the cross-validation is presented in Table 3.1. We can establish that λ_1=10 and λ_2=1 are optimal values of the weight parameters for the deterministic algorithm when it comes to the performance in the cross-validation. We can also make two other noteworthy observations in the table. First, for λ's where the ratio between λ_1 and λ_2 is 10, that is, along the diagonal in the table, the performance is similar. This observation implies that we need to allow for larger derivatives in the horizontal directions than in the vertical direction, which is quite natural since the gel tracks are clearly separated in the horizontal direction in multi-track gels. Second, for fixed λ_2, the performance for low values of λ_1 are very similar. This observation implies that the performance of the algorithm is not sensitive to how much we penalise derivatives in the vertical direction. However, this is probably a bit misleading because for low values of λ_1, the major control of large derivatives in the vertical direction lies in that we have restricted the change in this direction to lie between zero and two pixel widths.

$E \times 10^{-3}$		λ_2				
		0.001	0.01	0.1	**1**	10
	.01	190	166	155	145	
	.1	186	162	154	145	194
λ_1	1	199	161	151	145	193
	10	400	175	152	**142**	195
	100	863	402	181	147	205
	1000			470	238	241

Table 3.1: Cross-validation results for different values of the weight parameters. The tabulated E is the sum of the cross-validation errors for the three gels. The optimal values are highlighted in **boldface**.

Figure 3.2 shows the result of a simultaneous unwarping and classification of the three gels using the deterministic algorithm with the optimal weight parameters λ_1=10 and λ_2=1. Another visualisation of the result is to reorder the unwarped tracks and group them by class. This is done in Fig. 3.3.

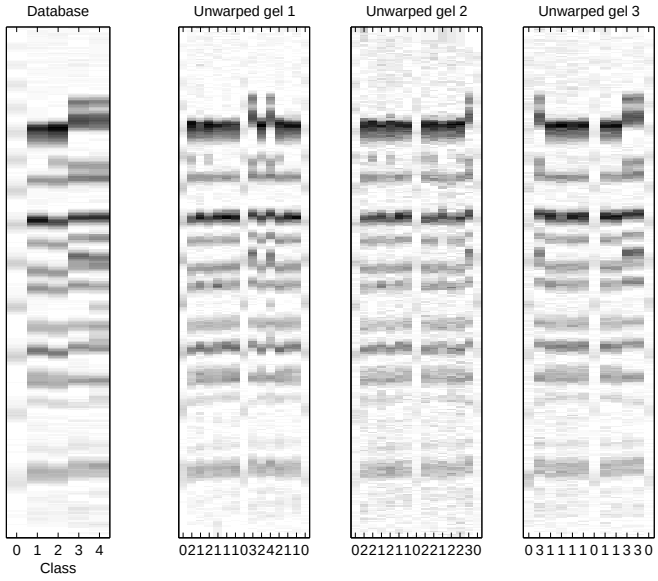


Figure 3.2: Database and unwarped and classified tracks from the three gels. The classification of the tracks is shown below each gel.

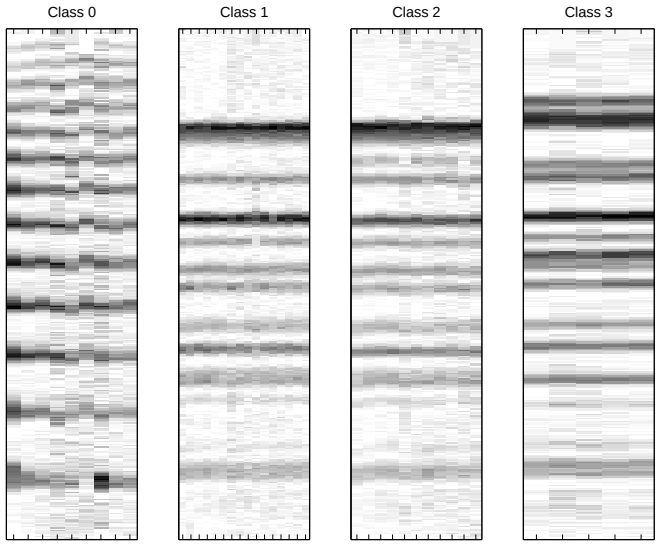


Figure 3.3: Unwarped and classified tracks from the three gels, reordered and grouped by class. Class 4 is not included in the figure since there is only one track classified to that class.

The result of the deterministic algorithm as shown in Fig. 3.2 and 3.3 is satisfying in a visual inspection but we cannot know for certain weather the algorithm has found the globally optimal solution or not, that is, the global minimum of S in Eq. (3.1). We therefore proceed by investigating if the performance can be improved using the stochastic algorithm.

Figure 3.4 shows the obtained value of S when increasing the number of iterations in the cooling schedule in the stochastic algorithm, and as comparison the obtained value of S for the deterministic value is marked with a dotted, horizontal line. We can conclude that for gel 1 and 3, the deterministic algorithm indeed does not find the globally optimal solution since the stochastic algorithm find a solution with even smaller S when the cooling is done slowly enough (a number of iterations larger than 100 seems to be enough). However, the performance of the stochastic algorithm is puzzlingly poor on gel 2. We interpret the two clusters in Fig. 3.4 (b) that the algorithm gets stuck in or near two deep local sub-optima of S for this gel. Further research needs to be done on this issue.

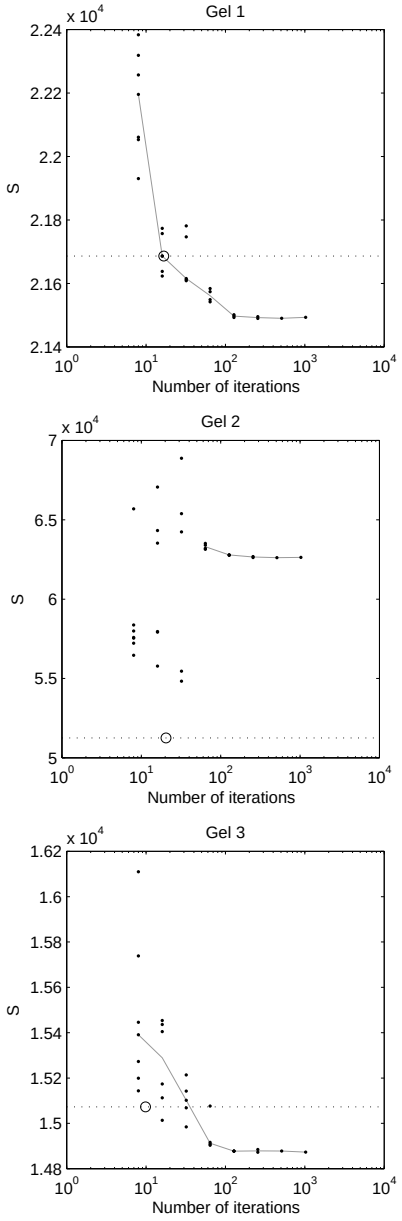


Figure 3.4: Comparison of the two algorithms. The value of S is plotted against the number of iterations used. The result of the deterministic algorithm is marked with a circle. The values for one or several runs/realisations of the pseudo-SA algorithm are marked with dots for a number of different cooling schedules. A median trace of the results for the pseudo-SA is shown with a grey, solid line. For comparison, the obtained value of S for the deterministic algorithm is marked with a dotted, horizontal line. ($\lambda_1 = 10$ and $\lambda_2 = 1$)

3.2 On two generalisations of the logarithmic data transformation

A useful concept in comparisons of different data transformations is the strength of a transformation. This complementary study uses that concept to compare two generalisation of the logarithmic data transformation in a situation where intuition fails. These two generalisations were investigated in the work that resulted in Paper III, and this study is closely related to the discussion in the second paragraph of Section 3.1 of Paper III.

Background

Box and Cox (1964) formulated a family of power transformations parameterised by the power λ_1 and offset λ_2. The transformations in the family are

$$h_\lambda(z) = \begin{cases} \dfrac{(z+\lambda_2)^{\lambda_1} - 1}{\lambda_1} & (\lambda_1 \neq 0) \\ \log(z+\lambda_2) & (\lambda_1 = 0) \end{cases} \tag{3.2}$$

and hold for $z > -\lambda_2$. This construction ensures that the transformations are continuous with respect to both the arguments and the two parameters. The power transformation family can be seen as a generalisation of the logarithmic transformation if we start with both parameters set to zero and then change to non-zero values of one or both of the parameters.

Another generalisation of the logarithmic transformation is the inverse hyperbolic sine transformation family formulated independently by Huber et al. (2002) and Durbin et al. (2002) in the context of variance-versus-mean stabilisation of microarray data. The family is motivated by a two-component model of measurement error where the measured value, denoted Z, is modelled as (Rocke and Lorenzato, 1995)

$$Z = \zeta e^{\eta} + \varepsilon + \alpha \tag{3.3}$$

Here ζ is an idealised value of the measured quantity, α is an additive bias, and ε and η are additive and multiplicative random error components, respectively. The two random error components are assumed to be normally distributed with expected value zero and variances σ_η^2 and σ_ε^2, respectively. The variance-versus-mean dependence of data distributed according to model (3.3) is stabilised approximately by a transformation that

is based on the inverse hyperbolic sine function

$$h_\lambda(z) = \text{arsinh}\left(\frac{z+\lambda_2}{\lambda_3}\right) \tag{3.4}$$

This gives us a transformation family that is parameterised by λ_2 and λ_3 with the restriction that $\lambda_3 > 0$. These parameters are related to the parameters in the model by $\lambda_2 = -\alpha$ and $\lambda_3 \approx \sigma_\varepsilon/\sigma_\eta$.

Given data, the transformation parameters that stabilise the variance-versus-mean dependence approximately can be estimated with the method used by Box and Cox (1964). Let z_{ijk} denote the measured value for protein i, treatment j and replication k. To obtain a measure of variance-versus-mean stabilisation, we compute the log-likelihood of original data under the assumption that for some unknown λ the transformed data is normally distributed with common variance, $h_\lambda(z_{ijk}) \sim N(\mu_{ij}, \sigma^2)$. For fixed λ, the maximised log likelihood with respect to the parameters μ_{ij} and σ^2 is

$$L_{max}(\lambda) = -\frac{1}{2}m(\log(2\pi)+1) - \frac{1}{2}m\log\hat{\sigma}^2(\lambda) + \sum_{ijk}\log\left|\frac{\partial h_\lambda(z_{ijk})}{\partial z_{ijk}}\right| \tag{3.5}$$

Here m is the total number of spots and $\hat{\sigma}^2(\lambda) = \sum_{ijk}(y_{ijk} - \bar{y}_{ij\cdot})^2/m$ is the maximum likelihood estimate of σ^2 for fixed λ.

Figure 3.5 shows contours of $L_{max}(\lambda)$ for an artificial data set that mimics the quantitative 2-D gel data in the 2×5 gel set in Paper III (842 proteins, two treatments and five replicates of each treatment) but where the measurement errors are simulated according to the two-component model in equation (3.3). The value of ζ for the proteins is taken from the 2-D gel data, whereas the other parameters are set to $\alpha = -50$, $\sigma_\eta = 0.01$ and $\sigma_\varepsilon = 2$. Hence, the proper parameters in an arsinh-transformation are $\lambda_2 = 50$ and $\lambda_3 = 200$. We can see that the maximum of $L_{max}(\lambda)$ in Figure 3.5(b) is close to these values.

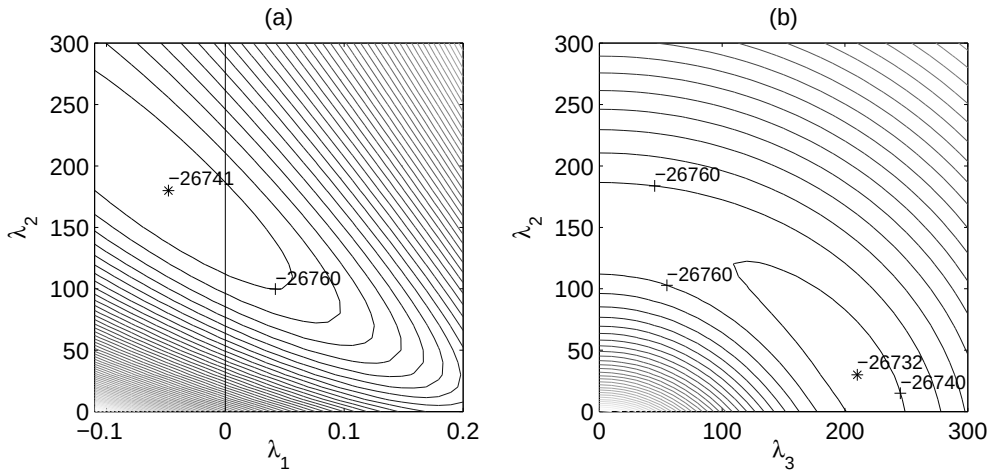


Figure 3.5: Contours of $L_{max}(\lambda)$ for (a) power transformation and (b) inverse hyperbolic sine transformation. The step between consecutive contours is 20 units in the log likelihood. The locations of maxima are marked by stars and the corresponding maximal values are shown.

Question at issue

The inverse hyperbolic sine transformation is never stronger than the logarithmic transformation, and a power transformation with $\lambda_1 < 0$ is intuitively stronger than the logarithm. Now, the maximum in Fig. 3.5(b) occurs for $\lambda_3 > 0$, indicating a transformation strictly weaker than the logarithm. Then, how can the maximum in Fig. 3.5(a) occur for $\lambda_1 < 0$?

We answer the question above using the concept of power strength of a transformation which was introduced by Tukey (1957). It can be defined as

$$k(z) = -\frac{\mathrm{d}^2 h/\mathrm{d}z^2}{\mathrm{d}h/\mathrm{d}z} z \tag{3.6}$$

This is not the exact definition made by Tukey (1957), but it can be derived from his formulation. Further, to avoid confusion when referring to the power transformations, we will from now on use the term strength instead of power strength.

For the power transformation family the strength is

$$k_\lambda(z) = (1 - \lambda_1)\frac{z}{z + \lambda_2} \tag{3.7}$$

and for the arsinh-transformation family the strength is

$$k_\lambda(z) = \frac{z(z+\lambda_2)}{(z+\lambda_2)^2 + \lambda_3^2} \tag{3.8}$$

We can see that the strength of the power transformations is $1 - \lambda_1$ when $\lambda_2 = 0$. For example, the square root transformation ($\lambda_1 = 1/2$) has strength $1/2$ and the logarithmic transformation ($\lambda_1 = 0$) has strength one.

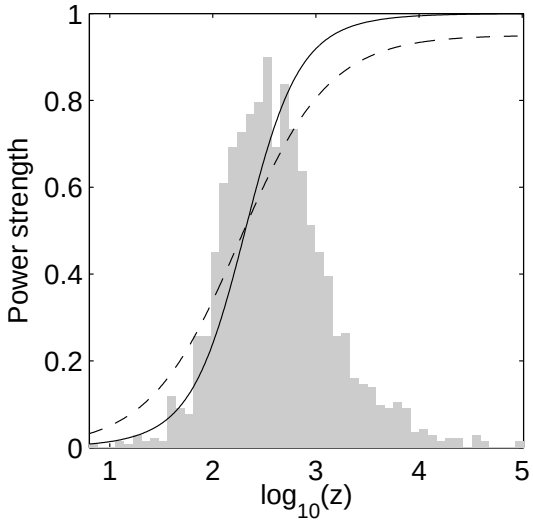


Figure 3.6: The strength as a function of $z$for the power transformation with $\lambda_1 = -0.05$ and $\lambda_2 = 180$ (dashed line), and for the arsinh-transformation with $\lambda_2 = 50$ and $\lambda_3 = 200$ (solid line). The histogram in grey shows the distribution of protein expression levels in the data.

To answer the question above we compare the strength of the power transformation corresponding to the maximum in Fig. 3.5(a) with the strength of the arsinh-transformation that corresponds to the maximum in Fig. 3.5(b). In figure 3.6 the strengths of these two transformations are plotted as a function of z. The solid line shows the strength of the arsinh-transformation with $\lambda_2 = 50$ and $\lambda_3 = 200$ and the dashed line the strength of the power-transformation with $\lambda_1 = -0.05$ and $\lambda_2 = 180$.

We can make two important observations about the strength of the power transformation. First, the parameters are tuned such that, for values of z in the bulk of the distribution of the data, the strength of the transformation resembles the strength of the arsinh-transformation. Recall that the

arsinh-transformation is optimal for this artificial data set since the data is generated using the model in Eq. (3.3). Second, the strength of the power transformation is between zero and one over the range of z in the data and therefore weaker than the log-transformation which has strength one. Hence, we were mislead by the intuition that told us that $\lambda_1 < 0$ indicated a power transformation stronger than the logarithmic transformation. The reason why the intuition fails is because the offset λ_2 is added before taking the power in the power transformations. However, returning to Fig. 3.5, we can see that for each fixed value of λ_2, the intuition holds. When there is a maximum for $\lambda_1 < 0$ in (a), the maximum in (b) is for $\lambda_3 \approx 0$.

Bibliography

Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (1998). *Essential Cell Biology*, Garland, New York.

Blomberg, A., Blomberg, L., Norbeck, J., Fey, S., Larsen, P., Larsen, M., Roepstorff, P., Degand, H., Boutry, M., Posch, A. and Görg, A. (1995). Interlaboratory reproducibility of yeast protein patterns analyzed by immobilized pH gradient two-dimensional gel electrophoresis, *Electrophoresis* **16**: 1935–1945.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B* **26**: 211–246.

Burstin, J., Zivy, M., de Vienne, D. and Damerval, C. (1993). Analysis of scaling methods to minimize experimental variations in two-dimensional electrophoresis quantitative data: Application to the comparison of maize inbred lines, *Electrophoresis* **14**: 1067–1073.

Conradsen, K. and Pedersen, J. (1992). Analysis of two-dimensional electrophoretic gels, *Biometrics* **48**: 1273–1287.

Dowsey, A., Dunn, M. J. and Yang, G.-Z. (2003). The role of bioinformatics in two-dimensional gel electrophoresis, *Proteomics* **3**: 1567–1596.

Draper, N. R. and Smith, H. (1998). *Applied regression analysis*, John Wiley & Sons.

Durbin, B., Hardin, J. S., Hawkins, D. and Rocke, D. (2002). A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics* **18**: S105–S110.

Garrels, J. I. (1989). The QUEST system for quantitative analysis of two-dimensional gels, *Journal of Biological Chemistry* **264**: 5269–5289.

Glasbey, C. A. (2005). Image warping and segmentation using generalisations of dynamic programming, Talk at AMSI (Australian Mathematical Sciences Institute) Symposium on Recent Advances in Biostatistics, Bioinformatics and Markov Chain Monte Carlo, Sydney, Australia, July, 2005.

Glasbey, C. A. and Mardia, K. V. (1998). A review of image-warping methods, *Journal of Applied Statistics* **25**: 155–171.

Glasbey, C. A. and Mardia, K. V. (2001). A penalized likelihood approach to image warping, *Journal of the Royal Statistical Society, Series B* **63**: 465–514.

Görg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R. and Weiss, W. (2000). The current state of two-dimensional electrophoresis with immobilized pH gradients, *Electrophoresis* **21**: 1037–1053.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics* **18**: S96–S104.

Mahon, P. and Dupree, P. (2001). Quantitative and reproducible two-dimensional gel analysis using Phoretix 2D Full, *Electrophoresis* **22**: 2075–2085.

Molloy, M. P., Brzezinski, E. E., Hang, J., McDowell, M. T. and VanBogelen, R. A. (2003). Overcoming technical variation and biological variation in quantitative proteomics, *Proteomics* **3**: 1912–1919.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins, *Journal of Biological Chemistry* **250**: 4007–4021.

Pandy, A. and Mann, M. (2000). Proteomics to study genes and genomes, *Nature* **405**: 837–856.

Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (eds) (2003). *The Analysis of Gene Expression Data*, Springer.

Pleissner, K.-P., Oswald, H. and Wegner, S. (2001). *Proteomics: From protein sequence to function*, BIOS Scientific Publishers, Oxford, chapter Image analysis of two-dimensional gels, pp. 131–149.

Rabilloud, T. (2002). Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but still climbs up the mountains, *Proteomics* **2**: 3–10.

Rocke, D. and Lorenzato, S. (1995). A two-component model for measurement error in analytical chemistry, *Technometrics* **37**(2): 176–184.

Scheffé, H. (1959). *The Analysis of Variance*, Wiley.

Smilansky, Z. (2001). Automatic registration for images of two-dimensional protein gels, *Electrophoresis* **22**: 1616–1626.

Speed, T. P. (ed.) (2003). *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall.

Tukey, J. W. (1957). On the comparative anatomy of transformations, *Annals of Mathematical Statistics* **28**: 602–632.

Vali, L., Wisely, K. A., Pearce, M. C., Turner, E. J., Knight, H. I., Smith, A. W. and Amyes, S. G. B. (2004). High-level genotypic variation and antibiotic sensitivity among *Escherichia coli* O157 strains isolated from two scottish beef cattle farms, *Appl. Environ. Microbiol.* **70**(10): 5947–5954.

Veeser, S., Dunn, M. J. and Yang, G.-Z. (2001). Multiresolution image registration for two-dimensional gel electrophoresis, *Proteomics* **1**: 856–870.

Voss, T. and Haberl, P. (2000). Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: Consequences for comprehensive data analysis, *Electrophoresis* **21**: 3345–3350.

Appendix A

Biological terms

Based on the definition at the web site http://www.tnc.se/bioterm/index.html.

Amino acid
Organic compound with the general formula $^{3}H_3H$-CHR-C00^{-}.

DNA (deoxyribonucleic acid)
Nucleic acid which constitutes the *genome* among all known *organisms* and some *viruses*.

Genome
The *DNA* or, for some *virus* (RNA virus), the *RNA* that carries the complete genetic information of the cell or virus.

Genomics
Scientific area that comprises mapping, *sequencing* and other simultaneous analyses of the whole or large parts of *genome*.

Nucleic acid
Chain of *nucleotides* linked by phosphodiesther binding.

Peptide
Unbranched chain of two or more *amino acids* linked by peptide binding.

Protein
Three-dimensionally structured molecular units consisting of one or several *peptide* chains.

Proteome
Those of a cell synthesised *proteins* that exist in that cell at a certain moment.

Proteomics
Scientific area that comprises simultaneous study of the whole or large parts of the *pro-*

teome.

Paper I

Paper II

Electrophoresis 2002, **23**, 1731-1744

Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern

John S. Gustafsson, Anders Blomberg and Mats Rudemo

Errata

Please find below some corrections to the above mentioned paper.

Page	Column	Line	Reads	Should read
1733	2	Eq. (3)	$v = -\nu\varepsilon\nabla\Psi$	$v = -\nu\nabla\Psi$
1736	1	-1	[26]	[25]
1736	2	Eq. (16)	$\frac{\partial f_i}{\partial x_j \partial x_k}$	$\frac{\partial^2 f_i}{\partial x_j \partial x_k}$
1737	1	24	[28]	[27]
1743	1	-13	[26]	[25]

Paper III

Paper IV

Comparative proteomics of industrial lager yeast reveals strain dependent differential expression of the *cerevisiae* and non-*cerevisiae* parts of their genomes

Robert Caesar, Ellinor Pettersson, Amir Hashemi, John S. Gustafsson[1], Anders Blomberg*

Department of Cell and Molecular Biology, Microbiology, Göteborg University, Medicinaregatan 9c, Box 462, SE-405 30 Göteborg, Sweden

[1]Department of Mathematical Statistics, Chalmers University of Technology, SE-41 296 Göteborg, Sweden

* Corresponding author

anders.blomberg@gmm.gu.se

tel: +46 31 7732589

fax: +46 31 7732599

ABSTRACT

Proteome analysis of the three industrial lager beer strains CMBS33, OG2252 and A15 were conducted under standardised laboratory growth conditions. We found the protein content of these three lager strains to be qualitatively rather similar, while they differ substantially to the *Saccharomyces cerevisiae* strain BY4742. Protein spots in the two-dimensional electrophoresis pattern of the lager strains were subjected to tandem mass spectrometry based identification, indicating that the non-*cerevisiae* proteins were most closely related to the corresponding proteins in *S. bayanus.* For many proteins the regulation of these *bayanus*-like proteins and their *cerevisiae* counterparts varied in a strain dependent manner, e.g. the *bayanu*-like form of Tdh3p was roughly 20-fold more abundant than the *cerevisiae* form in the OG2252 strain. In general, expression of the *cerevisiae* and the *bayanus*-like protein variants was more similar in the strains A15 and CMBS33. The phospoproteome of the lager strains were characterised using the Pro-Q Diamond stain. We found four novel phospho-proteins, Rsp12p, Efb1p, Rsp5p and Leu1p, but no qualitative differences in phosphorylation between the lager strains. In addition, no difference in protein N-terminal acetylation status was observed, generally indicating protein modifications to be of minor importance for the performance of lager strains.

INTRODUCTION

Lager beers are produced by what is called bottom-fermenting yeast strains (lager brewing strains) at low temperature (8º – 15º C). The fermentation is followed by a low-temperature period of maturation ("lagering") often lasting a few weeks. Lager beer currently constitutes more than 90% of the global beer production. Beer may have a alcoholic content ranging from a few to more than ten percent by volume, and tolerance to high alcohol levels is an important characteristic of lager beer strains. The industrial fermentation can also present a rather harsh environment to the yeast; hence, industrial strains typically display a high level of stress resistance. In addition, the yeast strain used in production influences the flavour of the beer. Better knowledge of specific production strains and the molecular features that relate to these industrial scale demands is desirable in order to optimize the industrial processes.

Lager brewing yeasts are in many respects very different from other brewing yeasts. It is also now well established that they are polyploid species hybrids formed from *Saccharomyces cerevisiae* and a closely related *Saccharomyces* species (1). In cases where the gene content of lager strains has been investigated in more detail, e.g. *ILV1* (2), *HIS4* (3) and *MET2* (4), two divergent forms of the gene in question has mostly been found. One of these forms has been shown to exhibit sequence similarity (e.g. by DNA hybridisation) to the corresponding *S. cerevisiae* gene. In the case of the non-*cerevisiae ILV1* and the *cerevisiae ILV1* the encoded proteins are 86% identical. In general, the similarity at the protein level seems to be in the range 78 – 88% (1). Boucherie and co-workers found that in the two-dimensional electrophoresis pattern of lager beer strains a number of abundant proteins appeared as "double spots", supporting the hybrid nature of these strains (5). They also found that the lager strains displayed similar protein 2D patterns, however, no analysis of the specific quantitative differences between the strains was performed. With the recent publication of full genome sequences for a number of *Saccharomyces* species (6,7) the understanding of the evolutionary history of lager strains should be more easily deciphered, extending the potential of applying proteomics technologies in the analysis of this important and interesting class of yeasts.

In this study we have analysed the proteome of three different commercial strains of lager yeast and found numerous quantitative differences among them. In

particular, the protein content was in all three cases very different to a standard laboratory strain of S. *cerevisiae*. Our results strongly support the hybrid nature of these three lager strains. In particular, it is interesting to note that the expression interplay between the *cerevisiae* and the non-*cerevisiae* part of the genomes (we provide evidence that the latter is most similar to *S. bayanus*) is for some specific proteins rather different. The present initial characterisation of these industrial strains will be of value for future more detailed analyses of how the *cerevisiae* and the non-*cerevisiae* (*S. bayanus*) part of their genomes respond during industrial scale fermentation. An understanding of the molecular features of these strains may have important industrial implications for strain improvements.

MATERIALS AND METHODS

Strains, Media and Growth Conditions

The strains used in this study were the laboratory strain BY4742 (Mat *his3*Δ*1, leu2*Δ*0, ura3*Δ*0, lys*Δ*0*) and the three industrial lager strains OG2252, A15 and CMBS33. The OG2252 strain was supplied by professor M. Kielland-Brandt (Carlsberg Laboratory, Denmark) although this strain originates from Alfred Jorgensens Laboratory, Denmark. A15 is a bottom fermenting industrial brewer's yeast provided for the present work by J.Londesborough, VTT Biotechnology, Espoo, Finland (the full name of the strain is VTT-A-63015). CMBS33 is a commercial lager strain from the KULeuven-CMBS collection. (Centre for Malting and Brewing Sciences) obtained through Prof. J. Thevelein and Prof. F. Delvaux (Leuven, Belgium).

All strains were cultivated at 30°C in SD medium (0. 14% yeast nitrogen base (Difco), 2% (w/v) glucose, 0.5% ammonium sulfate and 1% succinic acid, pH 5.8) supplemented with 120mg/l of uracil, histidine, leucine and lysine. Cells were inoculated, cultivated, radiolabeled and harvested as previously described (8). Protein concentration was measured with trichloroacetic acid precipitation using a Lowry-based commercial kit (Sigma P-5656).

Two-Dimensional Electrophoresis

Two-Dimensional electrophoresis was performed using 18 cm long non-linear immobilized pH gradient strips covering pH 3-10 (Amersham Bioscience). Protein extracts with a total radioactivity of 2 x 10^6 dpm was loaded on each analytic gel, 1 mg of protein was loaded on each preparative gel and 200µg protein was loaded on each Sypro-Ruby and Pro-Q Diamond gel. Rehydration of pH strips, first dimension, second dimension and comassie blue staining were carried out as previously described (8) using a Multiphor II (Pharmacia Biotech) for running the first dimension and an Ettan DALT II (Amersham Bioscience) for running the second dimension. Image preparation and analysis of radiolabeled gels were carried out as described elsewhere (9). For qualitative analysis of gel images using pseudo colour overlay, images where first warped into alignment as described elsewhere ((10); only warping step II). A CMBS33 gel where used as reference image in the pairwise image alignments. Sypro Ruby and Pro-Q Diamond staining were performed according to manufacturers protocol (Molecular Probes).

Protein Trypsinization, Mass Spectrometry and Analysis of MS/MS data

Gel pieces were cut out and in-gel trypsinated as described elsewhere (8) and peptides were subsequently eluted in 8µl elution buffer containing 2% acetonitrile and 0.05% formic acid. Peptides were separated on a Finnigan Surveyor chromatography workstation. The liquid chromatography (LC) device was directly coupled to an Finnigan LCQ ion trap mass spectrometer (Finnigan Corp., San Jose, CA). Peptide separation and peptide mass analysis were carried out as previously been described (9).

The SEQUEST search algorithm was used to correlate experimental spectra to theoretical spectra. Spectra were derived from the yeast.fasta database from the National Center for Biotechnology Information (NCBI) or from a fusion between the yeast.fasta and four databases containing open reading frame translations from *S. paradoxus, S. mikatae, S. bayanus* and *S. castelli*. ORF translation databases were retrieved from the SGD ftp-server (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/) and are based on sequence information from Cliften *et al.* and Kellis *et al.* (6,7). For proteins where no complete amino acid sequence information was available non-*cerevisiae Saccharomyces* DNA sequences retrieved from GeneBank (NCBI) were

manually translated into peptide sequences and added to the database. Protein identifications based on tandem mass spectra correlating to at least two tryptic peptides were considered valid for identification. For single charged peptides, only spectra with a cross correlation to a tryptic peptide of 1.5 or more were accepted. The corresponding value for multiple charged peptides was 2.0. Only peptides with a ΔCn score larger than 0.1 were accepted (11).

Phylogenetic analysis of non-*cerevisiae* proteins was carried out by compiling identified peptides into consensus sequences corresponding to amino acid sequences of the brewery strain proteins. ClustalW (1.83) multiple alignments of the regions covered by the MS/MS analysis were subsequently performed between the brewery strain sequences and the corresponding regions of the five *Saccharomyces* species included in the assay. The total length of the sequence analyzed was 1237 amino acid residues. The hierarchical tree was constructed as a N-J tree within ClustalW using multiple sequence alignment data.

RESULTS

Global protein expression differences between lager and laboratory strains

Proteome analysis of the three industrial lager beer strains CMBS33, OG2252 and A15 were conducted and their global protein content under standardised laboratory growth conditions was compared to the laboratory *S. cerevisiae* strain BY4742. Cultures were grown in defined minimal medium with glucose as the sole carbon and energy source and proteins were *in vivo* labelled during mid-exponential growth phase by the addition of ^{35}S-methionine. Cells were harvested, protein extract prepared and an equal amount of radioactivity was loaded onto the first dimensional isoelectric focusing gels spanning a pH range of 3 to 10. After second dimension separation gels were scanned and the obtained protein spots in the 2D pattern quantified by image analysis (roughly 500 spots were identified and matched in each of the patterns; in cases of uncertainty 2D patterns obtained after mixing of samples helped to elucidate the correct positional matching between spots). For all samples three independent experimental replicates were analyzed.

A first visual overview of the generated 2D patterns of the three lager yeast strains indicated that the overall spot pattern adhered mostly to the one generated

from the laboratory strain BY4742. However, detailed manual matching made clear that for quite a number of proteins qualitative differences could be scored (presence/absence of specific protein spots). In the colour overlay display in Figure 1A the individual gel patterns have first been warped according to a methodology developed earlier by us (10) to generate positionally identical 2D patterns, using CMBS33 as a standard. The warped patterns were then colour coded so that all spots from the CMBS33 strain appear red and all the spots from the other strain in the comparison appear blue (blue is the complement colour to the red; thus, when protein spots overlap in position the spots become black in this display). It is easily seen that the lager strains (exemplified in Figure 1A with the CMBS33 strain) displayed quite a number of qualitative differences compared to the laboratory strain. However, the differential colour display of the lager strains (compared to the CMBS33 strain) indicated that these three lager strains are very similar with few qualitative differences.

These visually observed qualitative differences were also reflected in the quantitative data; we found 62 proteins (15%) to exhibit a substantial difference in expression when comparing the laboratory strain BY4742 to CMBS33 (21 proteins were up-regulated and 41 down-regulated more than 5-fold in the CMBS strain). Roughly the same number of differences were found when comparing the two other lager strains to the laboratory strain (data not shown). However, in the comparison between the lager strains only a small number of proteins displayed quantitative differences greater than 5-fold (5 proteins in the case of OG2252 and 3 in the case of A15, in relation to CMBS33).

In order to get an overall similarity measure between strains based on quantitative data (all expression changes were taken into account) hierarchical clustering was performed (12). The cluster clearly shows that the lager strains were much more similar to each other than to the laboratory strain (Fig. 1B). However, it should be noted that the lager strain OG2252 stands out as the one with the largest quantitative discrepancy among the industrial strains. We conclude that the protein content of these three lager strains are rather similar, irrespective of if qualitative or quantitative differences are considered, while they differ quite substantially in both respects to the laboratory strain BY4742.

The non-*cerevisiae* proteins in the lager strain 2D pattern are most similar to *S. bayanus* proteins

It has earlier been reported that protein spots for many of the more dominant proteins appear in two different but closely situated variants in 2D analysis of lager strains (5); a *cerevisiae* spot plus a neighbouring spot with about the same intensity that is not found in the laboratory strain. We found ample examples of this in our detailed analysis of the three lager strains CMBS33, OG2252 and A15. Figure 2 displays some of these double spots in three zoom-in regions of the gels for lager strain CMBS33. It is clear that compared to the BY4742 strain, the CMBS33 strain displayed additional spots with roughly equal expression levels close to many of the more dominant proteins. Corresponding double spots to the ones found for CMBS33 was also found in the other two industrial strains (data not shown).

To identify these additional spots in the lager strains, spots were subjected to mass spectrometry based identification. Protein spots were cut-out from preparative 2D gels and in-gel trypsinized. The generated peptides were eluted and subjected to tandem mass spectrometry analysis on an ion-trap type mass spectrometer. Initially, generated peptide mass fragmentation spectra were matched to the *S. cerevisiae* proteome database. For 24 spots with no counterparts in the 2D pattern of the laboratory strain protein identities were revealed by matching to the *S. cerevisiae* protein database. In all cases a *S. cerevisiae* identity for the spot was also found. To find out if these extra protein spots belonged to the non-*cerevisiae* part of the lager genome we extended the fragmentation database matching to include also the proteomes from some of the other sequenced genomes of the *sensu strictu Saccharomyces* class (*Saccharomyces paradoxus*, *Saccharomyces mikatae* and *Saccharomyces bayanus;* (6)) as well as one non-*sensu strictu* species, *Saccharomyces castellii* (7). In doing so, we obtained for each of the species a catalogue of the peptides that was found to correspond to the MS generated fragmentation pattern. The analysis including all the other *Saccharomyces* proteomes was initially conducted on 7 different non-*cerevisiae* proteins taken from the CMBS strain. Using the data on the number of amino acids being identical in each of the species we constructed a phylogenetic tree based on the non-*cerevisiae* spots (Fig. 3). It is clear from the MS data that the non-*cerevisiae* spots are clearly most similar to sequences found in *S. bayanus* (Fig. 3). It should be noted that despite the fact that only a total of 1300 amino acids were included in this phylogenetic analysis, the

overall pattern of relatedness among the included species follows closely phylogeny data obtained either from full genome comparison (6) or from comparing 18S-28S rRNA spacer regions (13). The analysis was then extended to include all the non-*cerevisiae* proteins identified in the 2D analysis and in all cases the fragmentation pattern matched more closely to *S. bayanus* than to *S. cerevisiae*. Thus, our data on the non-*cerevisiae* spots found in these lager strains (protein spots not found in the laboratory strain) strongly support that these industrial lager strains are hybrids between *S. cerevisiae* and *S. bayanus*, or at least stem from a species/strain highly related to the sequenced MCYC623 strain of *S. bayanus*.

Expression differences between the lager strains are found for both *cerevisiae*- and *bayanus*-like proteins

For a central portion of the 2D gels with a high density of protein spots we conducted a much more thorough mass spectrometric analysis to identify most of the resolved proteins (Fig. 4). We cut out 150 proteins in this region from preparative 2D gels of CMBS33 and from the laboratory BY4742 strain. For the *S. cerevisiae* spots it was apparent that most of them occupy the same position in the two patterns (names indicated in green). A notable exception was found for the Adh1p protein for which there is a dominant spot in BY4742 while this protein was not found at the corresponding position in the CMBS33 strain. On the contrary, in the lager strain we found this *cerevisiae* Adh1p spot at a more acidic position; in addition, the expression of this *S. cerevisiae* variant of Adh1p was strongly down-regulated in the lager strains. However, in the lager strains we also identified the protein encoded by the isogene *ADH4* (the substantial similarity between *cerevisiae* and *bayanus* for this protein makes proper identification of its origin impossible). Unfortunately, the Adh4 protein overlaps with some other rather dominant proteins in the 2D pattern like Fba1p, which makes proper quantification of this interesting form of alcohol dehydrogenase impossible. Also the major spots for Tdh3p (this is clearly the *S. cerevisiae* derived protein) alters its position slightly to the acidic side in the lager strain CMBS33 indicating minor sequence differences or changes in post-translational modifications.

From this more detailed analysis quite a number of changes in position were apparent in the lager strains, highlighting the importance of independent MS based identity confirmation when analyzing these industrial strains. It is clear that many of

the major proteins, such as Eno1p, Eno2p and Fba1p, can be easily recognized in the lager strain patterns. However, for minor spots the identity was much less clear and MS identification actually revealed a rather large positional variability; e.g. the migration of protein spots like Adh1p, Gln1p, Tdh3p and Cys4p were rather different in the laboratory- and the industrial strains. It is clear from this more detailed analysis of these lager strains that one can not rely on protein identifications obtained from the analysis of laboratory *S. cerevisiae* strains. In this zoom-in portion of the gel we identified 32 proteins with a clear *S. cerevisiae* origin and 19 proteins with *S. bayanus* origin. Thus, our estimate for the overall contribution of proteins from *S. cerevisiae* to these lager strains are 63%. Currently, we have identified by MS roughly 300 proteins from the three different lager yeast patterns, and information from all the strains is summarised into the identification of 104 proteins in the lager strains (provided as supplementary material, SFig. 1).

After the confirmatory identification of proteins in the lager pattern a more detailed analysis of the expression of different proteins could be conducted. Comparing the different lager strains there are quite a number of significant changes in expression and even in the case of the two strains being most similar, CMBS33 and A15 (Fig. 1B), in total 38 proteins displayed at least a 2-fold change (Table I). A notable example is the down-regulation of both the *cerevisiae* and the *bayanus*-like Pdc1p forms in the A15 strain. Some examples of coordinated regulation of the *cerevisiae* and the *bayanus*-like forms were also found in the comparison between CMBS33 and OG2252; in the cases of Pyk1p/sbPyk1p (the prefix sb indicates a S. bayanus origin) and Eft1p/sbEft1p both protein forms were, for example, more highly expressed in the OG2252 strain. However, as in the case of aldehyde dehydrogenase Ald6p, where the *cerevisiae* form was down-regulated whereas the *bayanus*-like form was up-regulated, we could also find interesting examples of protein pairs being oppositely regulated in the OG2252 strain.

In Figure 5 the expression of all 19 spots for which both a *cerevisiae* and a *bayanus*-like variant was found and where proper quantification could be performed is displayed. For comparison the expression of the *cerevisiae* protein in the laboratory strain is also shown. For the *cerevisiae* spots it is clear that the strain differed slightly and in most cases the glycolytic enzymes appeared most highly expressed in the laboratory strain. Some clear examples of this are Tdh3p and Adh1p. However, for the industrial strains both the *cerevisiae* and the *bayanus* spots should be taken into

account and thus the total amount of that particular enzyme would be estimated by summing up the different forms in the industrial strains. In doing so it became apparent that for most of the proteins/enzymes, e.g. Tdh3p, Fba1p, the industrial and laboratory strains did not substantially differ in total amount of that protein.

It also became apparent that for many proteins the level of co-regulation of the *cerevisiae* and the *bayanus*-like variants varied between the strains. One can note that in the OG2252 strain the *bayanus* form of Tdh3p was roughly 20-fold more abundant than the *cerevisiae* form, while in the other two lager strains the two variants of Tdh3p were expressed to about the same level. Another example of this was Pyk1p for which the *bayanus*-like form is rather highly expressed in OG2252, while the *cerevisiae* form was kept low; once again the two forms were expressed to about the same level in the other two lager strains. Plotting all the pairs (*cerevisiae* versus *bayanus*-like form) it becomes apparent that the co-regulation for most pairs is rather strict in the CMBS33 and the A15 strains, while in the OG2252 strains the two evolutionary variants of many proteins displayed rather different levels of expression (Fig. 5b). We find these different regulatory pattern of genes from the two genomes in these hybrids interesting and conclude that they most certainly underlay many of the feature differences between these production strains.

In addition, taking both forms of the stress induced protein Sti1p into account the total expression appears to be much higher in the lager strains as compared to the laboratory strain (Fig. 5A). A general trend of high expression of stress proteins in the lager strains, exemplified by the 9 members of heat-shock proteins identified in the 2D gels (Fig. 6), was also observed. In particular the OG2252 strain consistently expressed these proteins to higher levels. It can be hypothesised that this is a reflection of the low optimal growth temperature for these strains, where the 30ºC in this laboratory setting is experienced as a heat-shock.

Analysis of the phosphoproteome of the lager strains.

We also extended this initial characterization of the proteome of these industrial strains to include analysis of post-translational modifications. The phospho-proteome of the strains was analyzed by use of the phospo-protein specific stain Pro-Q Diamond (14). The assay was found to function to our satisfaction on some test-proteins known to contain various amount of phosphorylations and the specificity for the phosphorylated forms of the proteins was clearly much higher (compare to the

band intensity in the phosphatase treated samples) (Fig. 7A).. The stain was then applied to the analysis of the whole phospho-proteome resolved in the 2D pattern. In comparison to the general protein stain Sypro Ruby it can clearly be seen that the phospho-specific stain detected much less proteins. Since the background activity for the Pro-Q Diamond stain is not totally abolished for non-phosphorylated proteins (Fig. 7A), we normalized for the amount of protein in the spots (a dominant protein spot could otherwise be scored as phosphorylated just because of its high abundance). The signal from Pro-Q Diamond was thus normalized to the signal from Sypro Ruby (d/s ratio) and to the length of the protein, and the resulting ratio for proteins displayed in Figure 7D. As can be clearly seen most of the strongly stained proteins by Pro-Q Diamond turned out to have high d/s ratios, but for some proteins, like Tdh3p, the ratio came out low, indicating the rather high Pro-Q Diamond signal only to be a result of their high expression. The spots with a high Pro-Q Diamond/Sypro Ruby ratio were cut out for identification by tandem MS (identified spots indicated in Fig. 7C). Roughly half of these proteins have previously been reported phosphorylated. We also found some novel phospho-proteins, namely Rsp12p, Efb1p, Rsp5p and Leu1p. However, when comparing the Pro-Q Diamond pattern for the three different lager strains we could not see any qualitative differences between the strains.

The lager strains displayed similar N-terminal acetylation as in *S. cerevisiae*

Another type of frequent post-translational modification is N-terminal acetylation, which also in some cases have been shown to be functionally important (9). The loss of this modification can easily be identified in the 2D pattern by a shift in pI (15). In the case of the industrial strains in our study we can not exclude that some of the spots, even being found at the same position in the 2D gels, could be protein forms with altered N-terminal acetylation. We thus analyzed our MS spectra for the identification of acetylated N-terminal peptides. For 8 proteins we identified an acetylated N-terminal tryptic peptide (Table II). In all these cases the acetylated peptide sequence appear to follow the same sequence rules as earlier found for *S. cerevisiae* (16), indicating that differences in the properties of these industrial strains might not be found in mutated N-terminal acetylation transferase systems with altered substrate specificity.

DISCUSSION

Minor but distinct differences between the proteomes of lager strains

The three lager yeast strains CMBS33, A15 and OG2252 displayed qualitatively very similar protein content. Thus, our data is in line with an earlier report where it was indicated that the 2D pattern between different lager yeasts were similar (5). However, we here also report on quantitative differences between the analysed lager strains. In particular we find it interesting that many of these differences in expression was scored for both *cerevisiae* and *bayanus*-like proteins (Fig. 5). The OG2252 strain stands out as most diverging from the other two with its highly unique expression pattern for some proteins, e.g. Tdh3p, sbPgk1p and Eft1p. However, the functional importance of these differences in protein regulation during industrial brewing conditions can not be ascertained at this stage since: i) the conditions under which we here have analysed these strains are far from the industrial process. This was not the intention of this study; on the contrary, we wanted to compare these lager strains under standardised laboratory conditions to be able to compare to the BY4742 strain. Which of the expression differences that will prevail or be enhanced during industrial scale fermentation will be the focus of follow-up studies where the here developed information about the 2D pattern of these industrial lager strains will be instrumental. ii) differences in performance of these lager strains can also depend on slightly different properties of the encoded enzymes. This can be exemplified by the result from a recent report where the authors overexpressed two variants of the *ATF2* gene, encoding an alcohol aetyltransferase, originating either from the lager strain CMBS33 or the ale yeast CMBS212 (these two proteins exhibit only 2 amino acid differences). It was found that overexpression of the lager version of Atf2p yielded substantially higher amounts of ester production (17).

Protein modification differences appear to be of minor importance

Differences in performance between strains could also be due to alterations in protein modifications that might alter protein functionality. A large variety of protein modification have been shown to occur on eukaryotic proteins, where the most common modification is N-terminal acetylations (16). This wide-spread modification has in some cases been shown to be essential for protein functionality (9). 2D analysis can be used to identify novel substrates by scoring for pI changes in N-terminal acetyl

transferase mutants (15). Differences in 2D position between the laboratory strain and the lager strains (Fig. 1A) could thus be caused by modification differences. MS analysis, however, revealed that for none of the analysed proteins that are known to be, or not to be, acetylated in the laboratory strain could we detect a change in the acetylation status in the lager strains. Even for the *bayanus*-like proteins there were no difference in N-terminal acetylation compared to the *cerevisiae* spots, indicating that the *cerevisiae* and *bayanus*-like NAT modification systems appear to show overlap in substrate recognition. This is in line with reports indicating that the N-terminal recognition sequence in proteins are the same when comparing different eukaryotic organisms (16). Another important modification is protein phosphorylations. In order to analyse this type of modification on a global scale we applied a recently developed phospohorylation specific dye. However, for none of the lager strains could we detect a substantial qualitative difference in phosphorylation pattern. Thus, differences in modifications, at least for N-terminal acetylations and phosphorylations, appear not to play a major role in setting physiological differences between these strains.

The phylogenetic classification of lager strains

Lager brewing strains are currently classified to belong to the *S. pastorianus* group. They seem to result from hybridisation between *S. cerevisiae* and another *Saccharomyces* yeast (1,18). There has been some uncertainty in the taxonomy of this species because of unclear features of certain type strains: e.g. in a 2D analysis of type strains of *S. pastorianus* that were supposed to be identical the proteome analysis revealed clear differences between the two (5). The relation to *S. bayanus* has also been unclear since some of its type strains appear themselves to be species hybrids, i.e. the type strain CBS380 of *S. bayanus* appears to contain two versions of many genes (1). It has more recently been indicated that *S. bayanus* (isolates IFO0359 and IFO1948) might represent lines which correspond to the non-*cerevisiae* genome of lager yeast (referred to as unpublished data in (1)). We here found that the non-*cerevisiae* protein spots in the 2D pattern were most closely related to the sequence from the recently sequenced *S. bayanus* strain MCYC623 (6). The lager yeast Weihenstephan 34/70 has been fully sequenced (referred to as unpublished data in (1)) and the average sequence identity to *S. bayanus* MCYC623 for the non-cerevisiae proteome is reported to be 92%. The other fully sequenced genomes *S. paradoxus* and *S. mikatae* of the *Sacharomyces sensu stricto* group display a much lower sequence

identity (84 and 82%, respectively) to the non-*cervisiae* part of the lager yeast genome. Thus, both full genome sequence analysis of the Weihenstephan 34/70 and the here reported MS analysis of non-*cerevisiae* proteins of the three commercial lager strains CMBS33, A15 and OG2252 support the close evolutionary relationship of the non-*cerevisiae* part of the lager genomes to the strain MCYC623 of *S. bayanus*.

The *cerevisiae* and *bayanus*-like proteins are differentially regulated in the lager strains

It now seems that hybrids and aneuploidy represent a rather common phenomenon in the *Saccharomyces* genus, especially among species and isolates used for industrial purposes (19). The minimum total size of the lager brewing yeast genome was found to be 23.3 million bp in the recently completed sequencing of the lager strain Weihenstephan 34/70 (unpublished data, cited in (1)). This genome size is approximately twice the size of the *cerevisiae* genome, and the number of non-*cerevisiae* genes is predicted to be more than 5294 (to be compared to the roughly 6193 *cerevisiae* genes in lager strains; i.e. 54% *S. cerevisiae* genes). In an earlier 2D based analysis of the qualitative pattern similarities between *S. cerevisiae* and the lager yeast k11 a somewhat larger number of *S. cerevisiae* proteins in the lager strain was indicated (83 - 89%) (5). Our detailed analysis of a smaller part of the 2D pattern (Fig. 4) indicated a *S. cerevisiae* contribution of proteins of 63%, more in line with the sequence-based data.

These *Saccharomyces* hybrids experience a highly intricate challenge in the proper regulation of the expression of genes encoded by the coexisting two sets of genomes. From our analysis we conclude that for most of the proteins a rather strict balance between the expression levels exists. For many of the genes the *cerevisiae* and the *bayanus*-like proteins are equally strongly expressed, in particular for the CMBS and A15 lager strains. However, in the case of OG2252 the situation was quite different. For a number of proteins we found quite divergent expression between the two gene sets, with Tdh3p and Ald6p being two extreme cases. Too little is known about the differences in biochemical features (kinetics, affinity, stability etc) when comparing *cerevisiae* and *bayanus* protein variants. Thus, whether the recorded expression of the *cerevisae* and *bayanus*-like variants is a prerequisite for proper performance of these strains in the industrial scale fermentation can not at this stage be determined.

The availability of the whole genome sequence and the here produced 2D map will enable the scientific community to carry out comprehensive expression analyses and genome structural analyses of lager brewing yeast strains. Application of such analyses to quality control in beer production and development of new products will be highly beneficial to the brewing industry. In addition, the unique protein features revealed by quantitative 2D analysis indicate that electrophoresis-based analysis could be used for patent protection to secure the uniqueness of specific industrial lager strains. The dynamics of *cerevisiae* and *bayanus*-like proteins during industrial scale fermentation and stress conditions will be highly interesting topics for future investigation. However, the complexity of the 2D pattern that were revealed by MS analysis of the lager strains indicated that proteomics analysis of these industrial strain has to based on thorough parallell MS identification, and can not be based on matching to previously identified 2D patterns from *S. cerevisiae* laboratory strains.

ACKNOWLEDGEMENT

The financial support came from the EC project “High-gravity brewing”. QLK1-CT2001-01066. JSGs work was financially supported by the Chalmers Bioscience programme. We also would like to express our sincere appreciation to constructive comments and suggestions from the “High-gravity brewing” consortia members during the course of this study, and to departmental colleagues for suggestions for improvements of the manuscript.

FIGURE LEGENDS

FIGURE 1. **Comparison of protein expression in BY4742, CMBS33, OG2252 and A15.** (A) a colour overlay display of warped images from the different lager strains and the laboratory strain. CMBS33 was used as the template in the warping step. In each of the comparisons the image/spots of the CMBS33 strain is coloured red while the image/spots from the other strain is coloured blue. In cases where the two spot patterns overlap the spot will appear black with the complementary colours chosen (B) phylogenetic tree of the four different yeast strains based on the positional (same x- and y- position) matching of the 2D gels. The tree is based on hierarchical clustering of expression data (12).

FIGURE 2. **Examples of double spots in the CMBS33 strain for three zoom-in regions of the whole 2D image.**

FIGURE 3. **Origin of non-*cereviseae* proteins in CMBS33.** Seven dominant spots on a CMBS33 gel lacking corresponding spots on BY 4742 gels were identified as Eft1p, Shm2p, Ilv5p, Pdc1p, Met6p, Pgk1p and Pyk1p using ESI-MS/MS. The SEQUEST algorithm was subsequently used to correlate MS/MS spectra from these spots to theoretical mass data derived from a FASTA database containing the amino acid sequences of the identified proteins and their homologs in *S. paradoxus, S. mikatae, S. bayanus* and *S. castelli*. For each protein the peptide sequences identified were compiled to a sequence for the CMBS33 form of the protein. In total, the peptides identified contained 1237 amino acids covering 32% of the seven proteins. Multiple alignment of the regions covered by the MS/MS analysis were performed between the five *Saccharomyces* species and the brewery strain using ClustalW and a hierarchical tree based on the result was constructed. The brewery strain differed with one amino acid from *S. bayanus*, with 36 amino acids from *S. cereviseae*, with 34 amino acids from *S. paradoxus*, with 32 amino acids from *S. mikatae* and with 134 amino acids from *S. castelli*.

FIGURE 4. **Zoom-in on the central portion in the 2D pattern.** The protein names for spots that have been identified as corresponding to S. *cerevisiae* in green and spots most similar to *S. bayanus* in red. The spots with their name in blue correspond to

cerevisiae spots that have changes in their position in the lager strains compared to their 2D position in the laboratory strain. Black names indicate proteins that have been identified via MS analysis but for which no clear link to either *S. cerevisiae* or *S. bayanus* was found. The range of molecular weight and pI is roughly 30 – 65 kDa and 5.5 – 7.2, respectively.

FIGURE 5. **Proteins with dual forms in lager strains.** (A) 24 proteins where both a *S. cerevisiae* and a *S. bayanus* form were identified are displayed. For each protein the scale on the y-axis of the two diagrams are the same. Error bars indicate ± SD. (B) the expression of 19 *S. bayanus* proteins plotted against the expression of the corresponding *S. cerevisiae* proteins. Relative expression is given as parts per million (ppm). Both axes are on logarithmic scale.

FIGURE 6. **Expression data for heat-shock proteins.** Error bars indicate ± SD.

FIGURE 7. **Analysis of the phosphoproteome of the yeast strains by use of the phospoho-specific stain Pro-Q Diamond.** (A) test of staining conditions using standard proteins know to be phosphorylated. For one lane is displayed the result after phosphatase treatment of the protein mixture. (B) the Sypro Ruby staining of total protein for strain CMBS33. (C) the Pro-Q Diamond staining of strain CMBS33. (D) the ratio between Pro-Q Diamond and Sypro Ruby staining was estimated for roughly 150 proteins. The ratio is also normalized for the length of the protein (longer proteins have stronger staining in Sypro Ruby). Numbers refer to the numbered arrows in Figure C.

Table 1. Differences in protein expression between the lager strains. The expression differences in A15 and OG was compared to CMBS33. Differences relate to statistical significance (individual $p<0.05$) and at least a 2-fold differences.

	# of proteins up-regulated	identified up-regulated proteins	# of proteins down-regulated	identified down-regulated proteins
CMBS vs A15	16	Eno2, Hom6	22	Pdc1, sbPdc1,
CMBS vs OG	43	Kar2, Eft1, sbEft1, Pyk1, sbPyk1, Ade17, Adh1, sbAdh1, sbFba1, sbAld6, sbTdh3	22	Ald6, Cys4, Pgk1, Tdh3

Table 2. Acetylation status of CMBS33 proteins (for information about Xcorr and Delta CN see Materials and Methods).

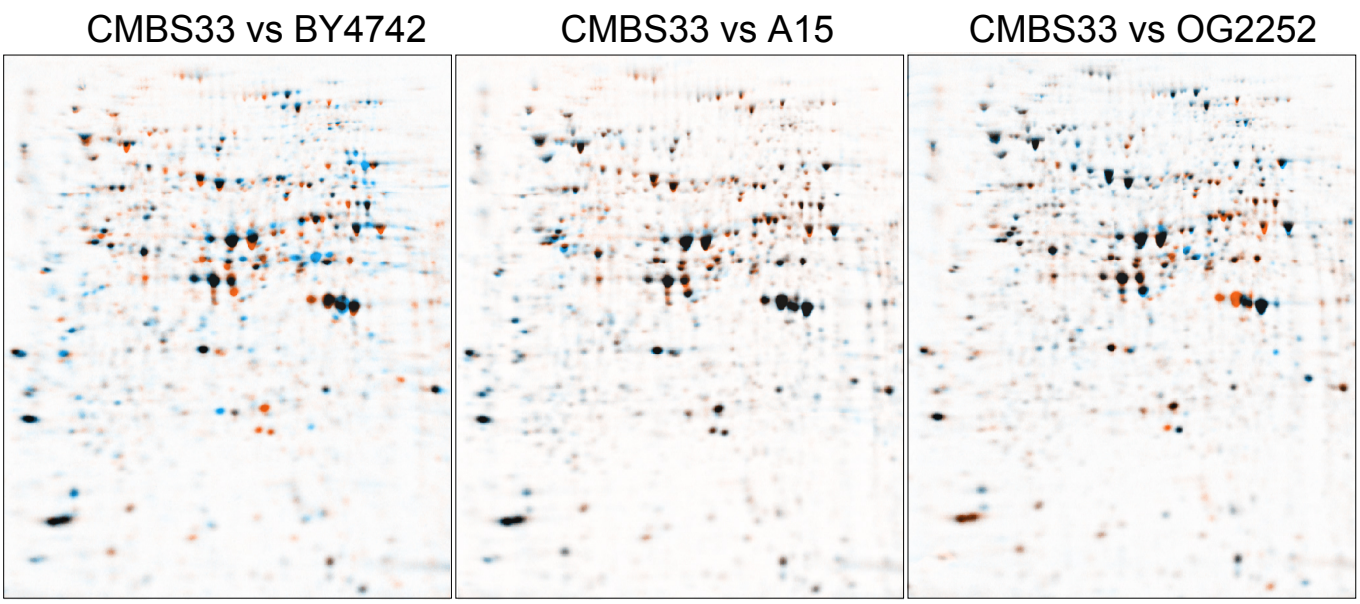
N-terminally acetylated proteins		MH+[a]	charge	Xcorr	Delta CN
Prc1	Ac-SEITL	789.86	1	1.63	0.63
Prc1 (*bayanus*)	Ac-SEITL	789.86	1	1.76	0.62
Gln1	Ac-AETSI	819.84	1	1.44	0.52
Pgk1	Ac-AFTFK	2345.60	3	0.57	0.33
Sti1	Ac-SLTAD	968.99	1	2.06	0.65
Sah1	Ac-SAPAQ	920.95	1	2.81	0.70
Efb1	Ac-ASTDF	797.80	1	2.63	0.71
Act1	Ac-MDSEV	2690.81	3	1.85	0.41
N-terminally unacetylated proteins					
Met6	VQSAV	1074.26	2	3.58	0.67
Eft1	VAFTV	1083.24	2	3.34	0.6
Eft1 (*bayanus*)	VAFTV	1083.24	2	1.64	0.54
Ilv5	MLRTQ	1720.97	3	1.3	0.09
Dld3	TAAHP	1668.88	3	4.66	0.61
Cys3	TLQES	1268.4	2	2.8	0.73
Ade13	PDYDN	1529.59	2	3.92	0.69
Fba1	GVEQI	786.94	2	2.53	0.67

a) the N-terminal peptides ion mass

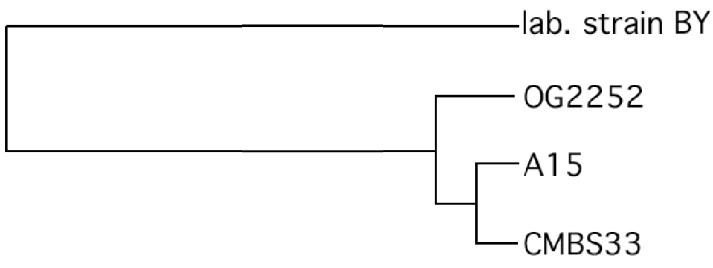
REFERENCES

1. Kodama, Y., Kielland-Brandt, M. C., and Hansen, J. (2005) in *Comparative Fungal Genomics* (Sunnerhagen, P., and Piskur, J., eds), pp. in press, Springer-Verlag, Heidelberg
2. Nilsson-Tillgren, T., Gjermansen, C., Holmberg, S., Petersen, J. G. L., and Kielland-Brandt, M. C. (1986) *Carlsberg Research Communications* **51**, 309-326
3. Holmberg, S. (1982) *Carlsberg Research Communications* **47**, 233-244
4. Hansen, J., and Kielland-Brandt, M. C. (1994) *Gene* **140**, 33-40
5. Joubert, R., Brignon, P., Lehmann, C., Monribot, C., Gendre, F., and Boucherie, H. (2000) *YEAST* **16**, 511-522.
6. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003) *Nature* **423**, 241-254
7. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M. (2003) *Science* **301**, 71-76
8. Blomberg, A. (2002) *Methods Enzymol* **350**, 559-584
9. Caesar, R., and Blomberg, A. (2004) *J Biol Chem* **279**, 38532-38543
10. Gustafsson, J. S., Blomberg, A., and Rudemo, M. (2002) *Electrophoresis* **23**, 1731-1744
11. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) *J. Am. Soc. Mass Spectrom.* **5**, 976-989
12. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) *Proc Natl Acad Sci U S A* **95**, 14863-14868.
13. Oda, Y., Yabuki, M., Tonomura, K., and Fukunaga, M. (1997) *YEAST* **13**, 1243-1250
14. Steinberg, T. H., Agnew, B. J., Gee, K. R., Leung, W. Y., Goodman, T., Schulenberg, B., Hendrickson, J., Beechem, J. M., Haugland, R. P., and Patton, W. F. (2003) *Proteomics* **3**, 1128-1144
15. Polevoda, B., Norbeck, J., Takakura, H., Blomberg, A., and Sherman, F. (1999) *EMBO J.* **18**, 6155-6168
16. Polevoda, B., and Sherman, F. (2003) *J Mol Biol* **325**, 595-622
17. Verstrepen, K. J., Van Laere, S. D., Vanderhaegen, B. M., Derdelinckx, G., Dufour, J. P., Pretorius, I. S., Winderickx, J., Thevelein, J. M., and Delvaux, F. R. (2003) *Appl Environ Microbiol* **69**, 5228-5237
18. Bond, U., and Blomberg, A. (2005) in *Food and Yeasts* (Querol, A., and Fleet, G., eds), Springer, Heidelberg
19. de Barros Lopes, M., Bellon, J. R., Shirley, N. J., and Ganter, P. F. (2002) *FEMS Yeast Research* **1**, 323-331

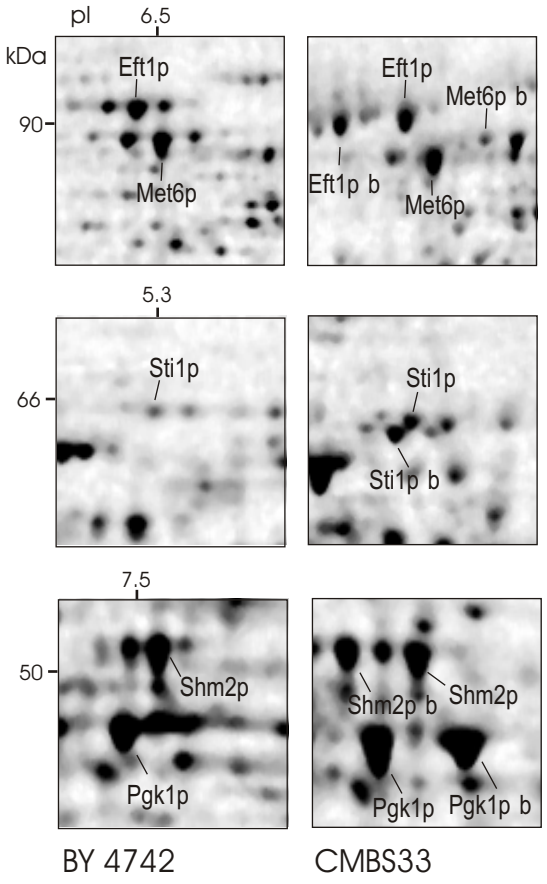
A.



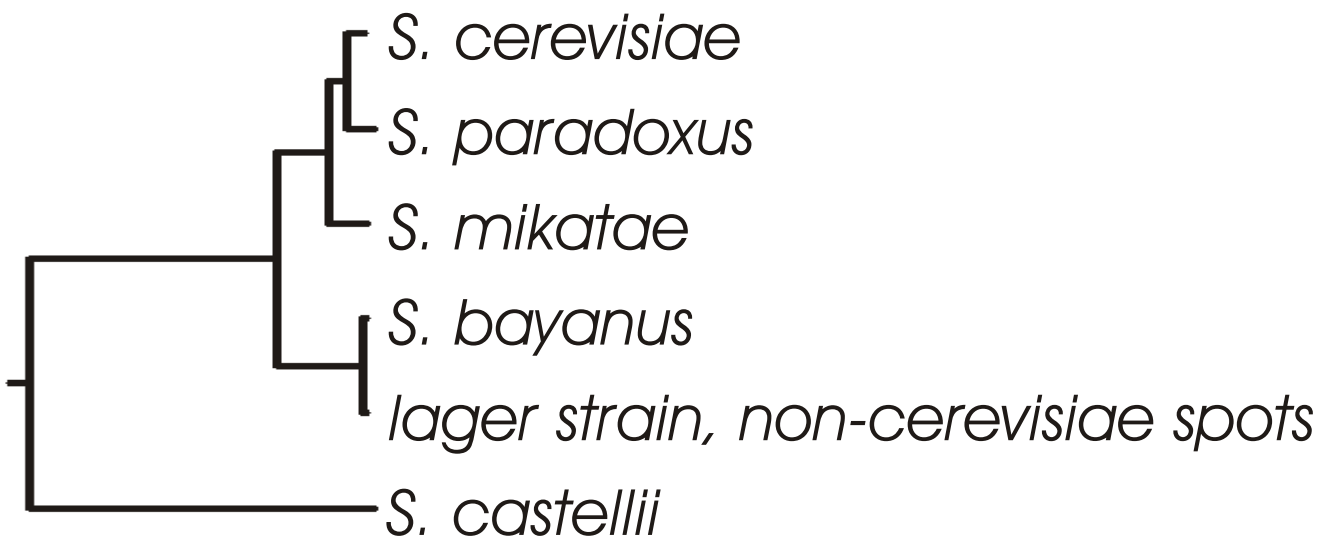
B.



Caesar *et al.* Figure 1

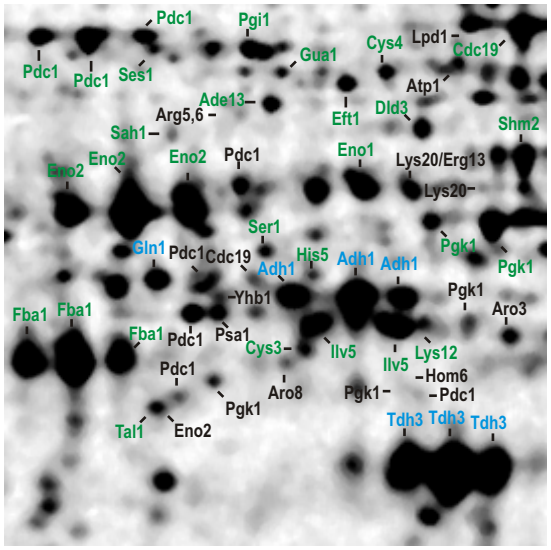


Caesar *et al.* Figure 2

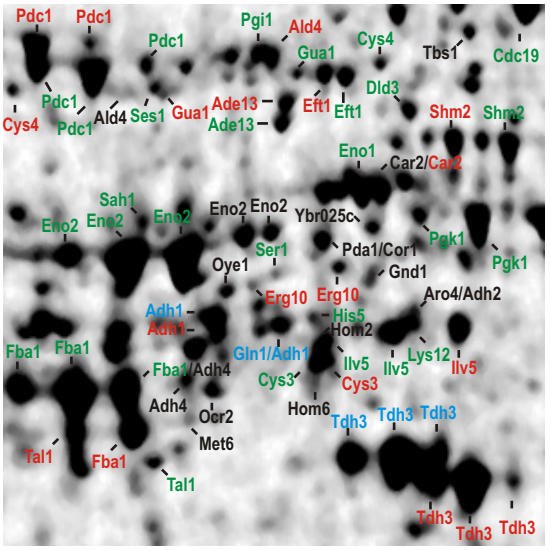


Caesar *et al.* Figure 3

BY 4742

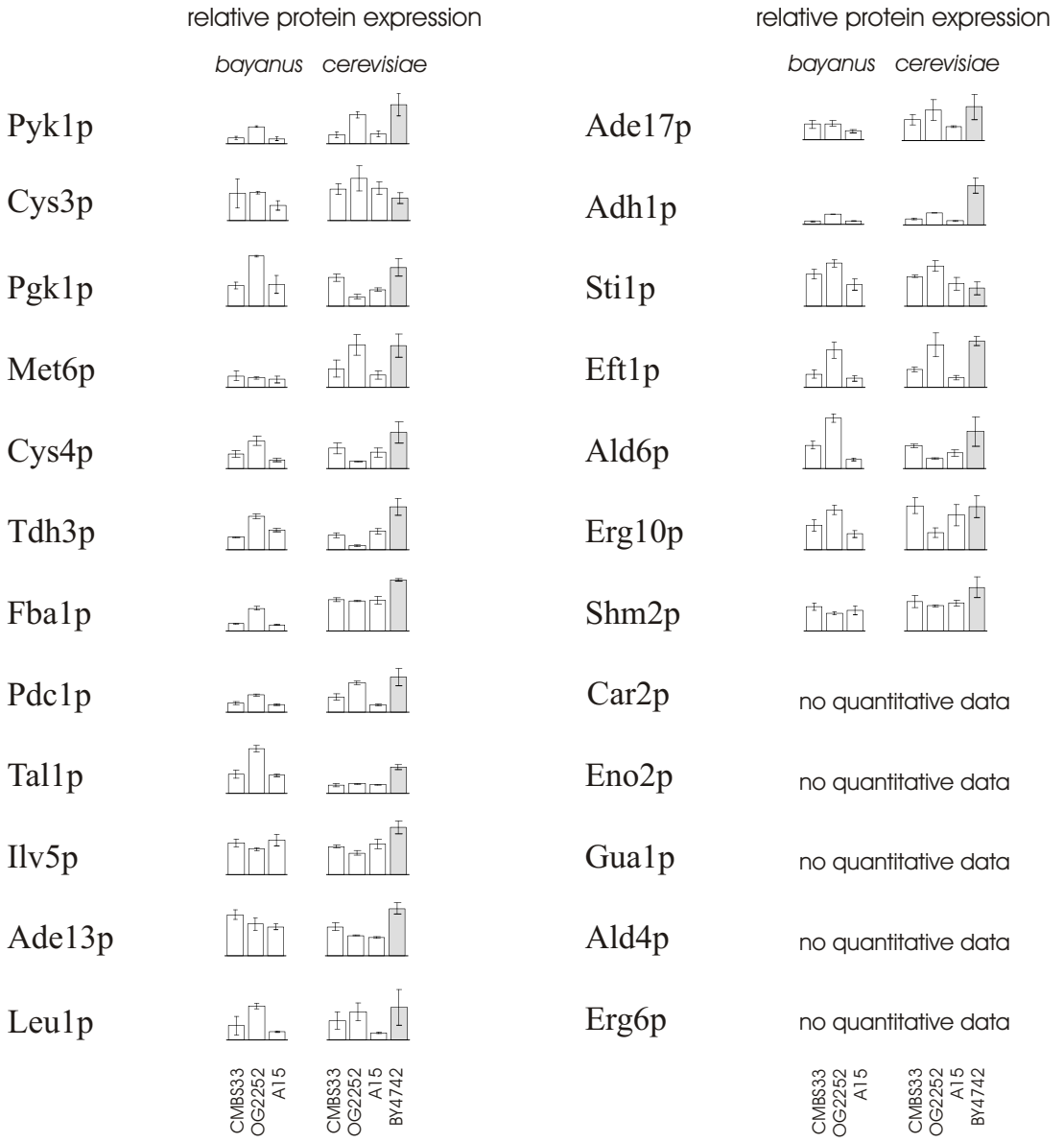


CMBS33

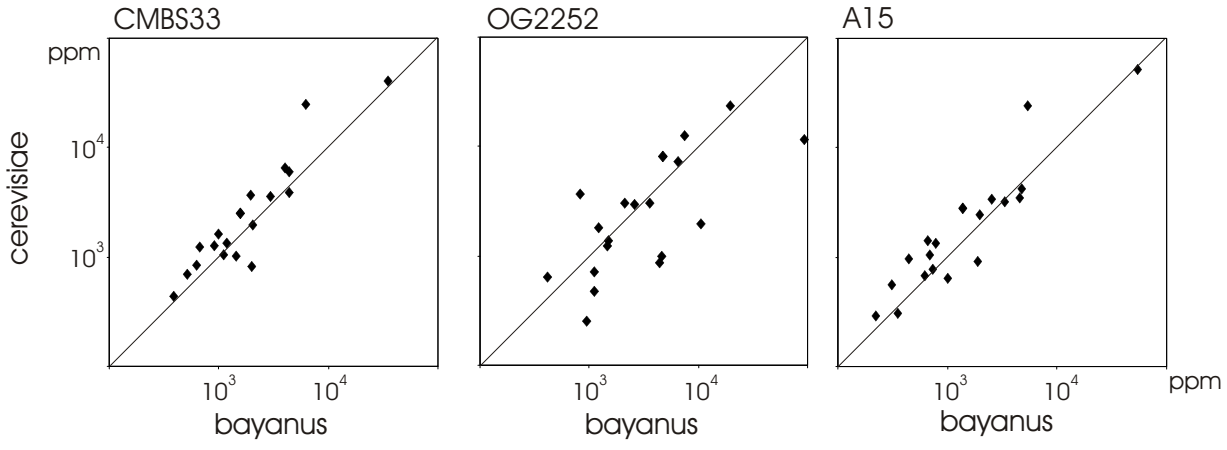


Caesar *et al.* Figure 4

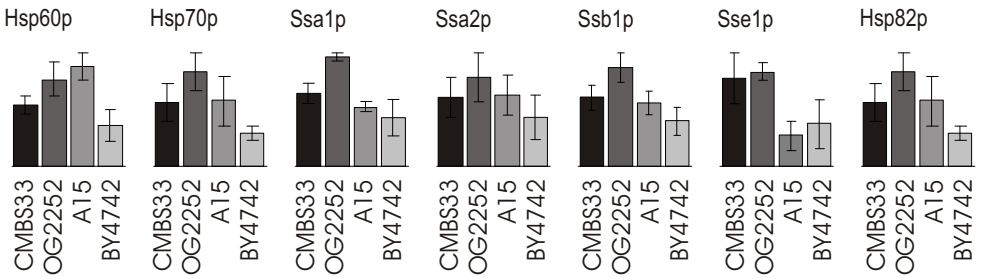
A.



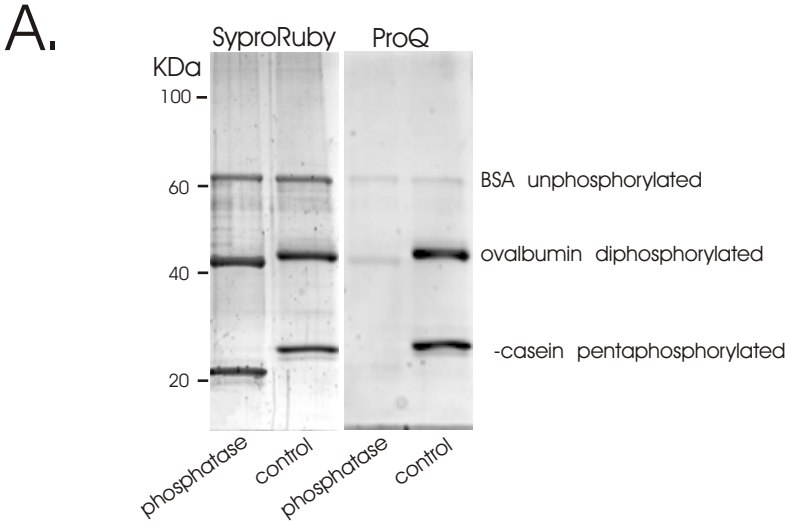
B.

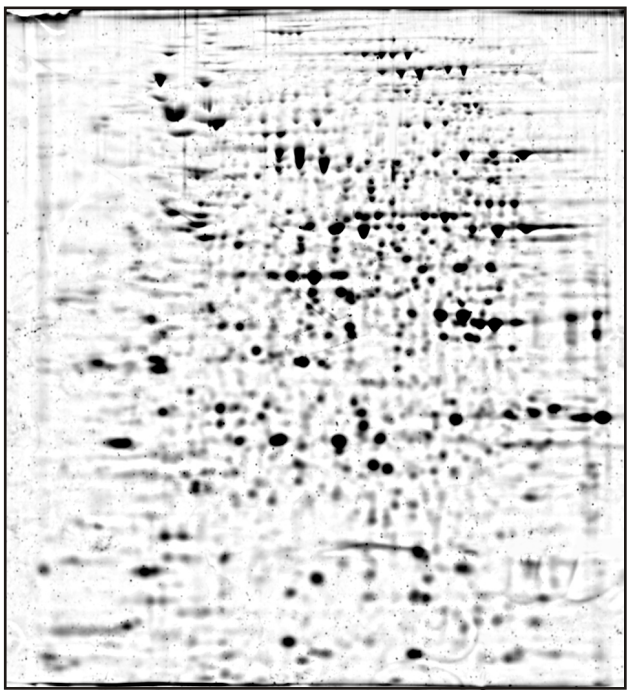


Caesar *et al.* Figure 5

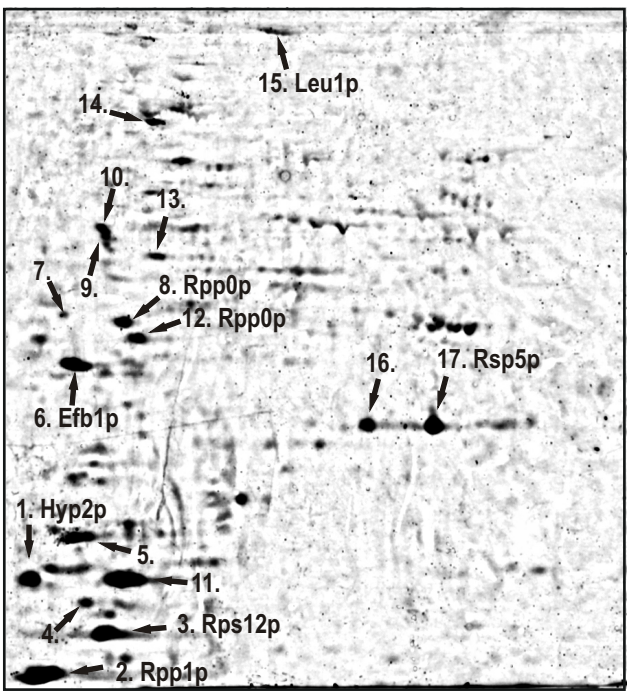


Caesar *et al.* Figure 6

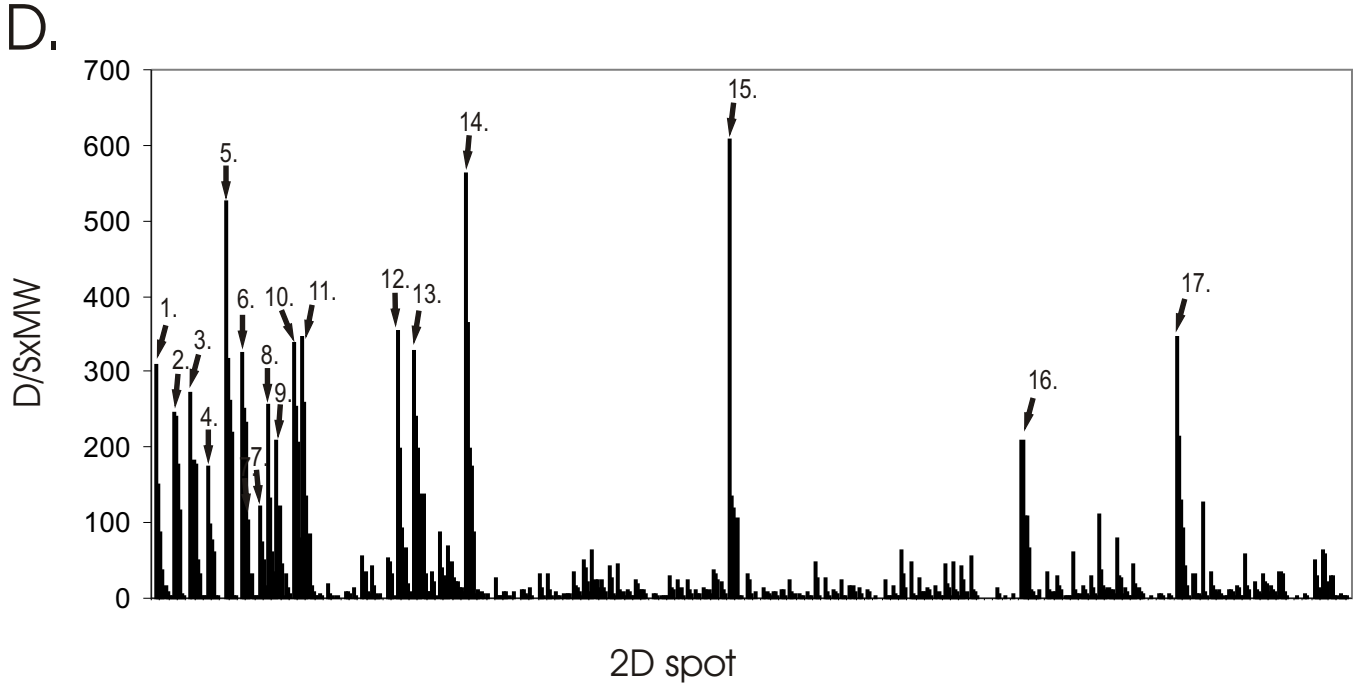




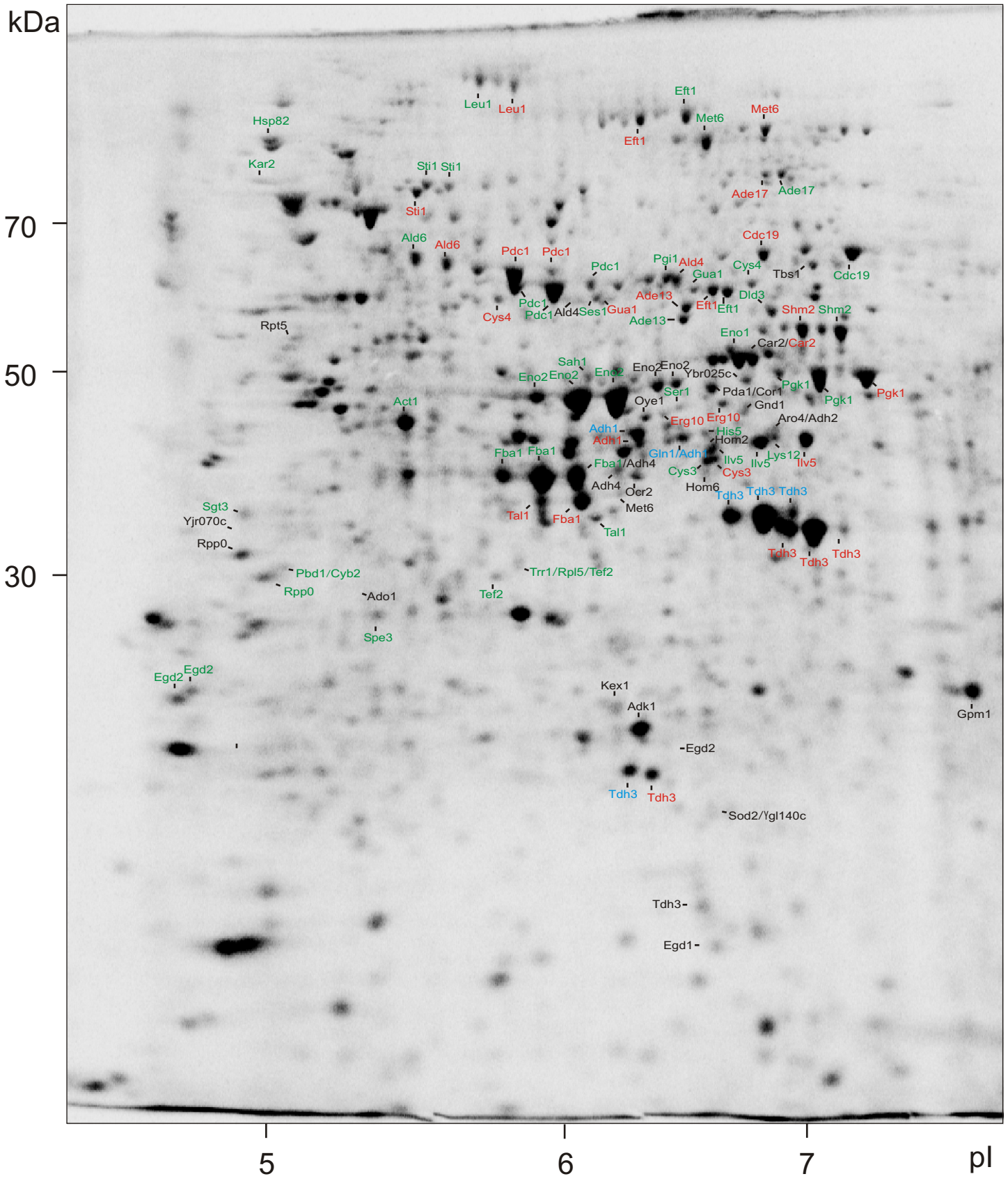
Sypro Ruby CMBS33



ProQ Diamond CMBS33



Caesar *et al.* Figure 7



Caesar *et al.* Suppelmentary inormation Figure 1