Workshop on

Statistical Aspects of Microarray Data

Thursday 20 - Saturday 22 February, 2003
Department of Mathematical Sciences
University of Aarhus

The lectures took place in Auditorium F, Building 534

Program, abstracts and list of participants

This booklet contains the abstracts of the talks at the Maphysto-workshop "Statistical Aspects of Microarray Data" that was held at the Department of Mathematical Sciences, University of Aarhus. The workshop was funded by MaPhySto and by a grant from the Danish National Science Foundation. The workshop was part of the MaPhySto initiative in mathematical modelling in biology and was organized by Jens Ledet Jensen (University of Aarhus), Mathisca de Gunst (EURANDOM, Eindhoven/Free University of Amsterdam), Mats Rudemo (Stochastic Centre, Gothenburg), and Michael Sørensen (University of Copenhagen).

It was the intention of the workshop to cover a number of the statistical issues of importance for the analysis of microarray data. This goes from the basic level of extracting information from the scanned images, defining a suitable expression level, and normalization issues when comparing different arrays, to the high level analysis involving clustering of genes and samples, finding differentiable expressed genes, and building classifiers for diagnostic purposes.

The booklet contains the programme for the workshop, the abstracts of the talks and a list of participants.

Contents

1 Program

Thursday, February 20

08.30–09.30	Registration
09.30–09.40	Welcome
09.40–10.25	Claus Mayer (Rowett Research Institute, Aberdeen, Scotland): Least trimmed squares methods for microarray normalization
10.40–11.10	Coffee/tea
11.10–11.55	David Edwards (NOVO Nordisk, Denmark): On the pre-analysis of one-channel cDNA microarray data
12.10–13.30	Lunch
13.30–14.15	Torben F. Ørntoft (Molecular Diagnostic Laboratory, Aarhus, Denmark): Microarrays, basic principle and use in medical research
14.30–15.15	Yudi Pawitan (Department of Medical Epidemiology, Karolinska Institute, Stockholm): Survival analysis using gene expression data
15.30–16.00	Coffee/tea
16.00–16.45	Mark van der Laan (University of California, Berkeley): Statistical inference with gene expression data

Friday, February 21

09.10–09.40	Mark Reimers (Center for Genomics and Bioinformatics, Karolinska Institute, Stockholm): Issues in probe level analysis of affymetrix data
09.50–10.20	Mette Langaas (Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway): Estimating the number of genes truly differentially expressed
10.30–11.00	Coffee/tea
11.00–11.45	Jelle Goeman (Leiden University Medical Center, The Netherlands): A global score test for differential expression of groups of genes in high-dimensional microarray data
12.00–13.30	Lunch
13.30–14.00	Jörg Assmus (Department of Mathematics, University of Bergen, Norway): On the problem of significant p-values for gene expressions
14.10–14.55	Mats Rudemo (Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden): Empirical Bayes analysis of variance models for microarray data
15.10–15.40	Coffee/tea
15.40–16.25	Ernst Wit (Department of Statistics, University of Glasgow, U.K.): Hidden Markov modelling of genomic expression interactions

16.40–17.25 Arnoldo Frigessi (Norwegian Computing Center, Oslo, Norway): Towards a comprehensive statistical model of cDNA microarrays

18.30: Workshop Dinner

Saturday, February 22

09.00–09.45 Rafael Irizarry (Department of Biostatistics, Baltimore, USA): Getting usable data from microarrays: The role of statisticians

10.00–10.30 Coffee/tea

10.30–11.15 Steen Knudsen (Technical University of Denmark): A new non-linear normalization method to reduce variablility in DNA microarray experiments

11.30–12.15 Volkmar Liebscher (Institute of Biomathematics and Biometry, München, Germany): Stochasting modelling and quality control for gene expression data

12.30–13.15 Lunch

13.15–14.00 Anja von Heydebreck (Max-Planck-Institute for Molecular Genetics, Division of Computational Molecular Biology, Berlin, Germany): Error modelling, data transformation and robust calibration for microarray data

14.15–14.45 Coffee/tea

2 Abstracts of lectures

Jörg Assmus, Hans Karlsen, and Dag Tjøstheim

University of Bergen, Norway

On the problem of significant p-values for gene expressions

It is well-known that the Bonferroni procedure is much too conservative to be of much help in identifiing significant genes. We will briefly review alternative procedures that have been proposed in the recent literature, and mainly introduce a new algorithm based upon non paramerically estimated p-value distributions. The algorithm is performed and tested on both simulated and real data.

David Edwards

Biostatistics Dept, Novo Nordisk, Denmark

On the pre-analysis of one-channel cDNA microarray data

Data from one-channel cDNA microarray studies may exhibit poor reproducibility due to spatial heterogeneity, non-linear array-to-array variation and problems in correcting for background. Uncorrected, these phenomena can give rise to misleading conclusions.

Spatial heterogeneity may be corrected by domain-specific global normalisation or by use of two-dimensional loess smoothing (Colantuoni et al., 2002). A method to correct non-linear between-array variation using an iterative application of one-dimensional loess smoothing is proposed. This method, called mean cyclic loess, is related to the cyclic loess method of Bolstad et al. (2002) but is more efficient. A simple method for background correction using a smoothing function rather than subtraction is described. These techniques promote within-array spatial uniformity and between-array reproducibility. Their application is illustrated using data from a study of the effects of an insulin sensitizer, rosiglitazone, on gene expression in white adipose tissue in diabetic db/db mice. The techniques may also be useful with data from two-channel cDNA microarrays and from oligonucleotide arrays.

References:

Bolstad, B.M., Irizzary, R.A., Astrand, M. and Speed, T.P. (2002). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. Submitted to Bioinformatics.
Colantuoni, C., Henry G., Zeger, S. and Pevsner, J. (2002). Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. Biotechniques, 32(6): 1316-20.
Edwards, D. (2003). Non-linear Normalization and Background Correction in One-Channel cDNA Microarray Data. Bioinformatics, to appear.

Arnoldo Frigessi and Ingrid K. Glad

Dept. of Mathematics, University of Oslo

Towards a comprehensive statistical model of cDNA microarrays

This is joint work with Heidi Lyng (Norwegian Cancer Hospital), Mark van de Wiel (Eindhoven University of Technology) and Marit Holden (Norwegian Computing Center). A microarray experiment is a complex sequence of several laboratory and computer related tasks, leading to an I * J matrix of (log ratio) gene expressions for I genes and J individuals or replications. Each such task is studied or modelled separately, and the 'result' of one step is plugged into the next step for further processing. This facilitates a good overview and optimal handling in each step, but creates also major drawbacks: uncertainty is not propagated through the sequence of various tasks, each step is based on a statistical model which might be contradictory from task to task, and it is not possible to perform joint inference from the experiment as a whole.

We present a coherent and structured statistical model which describes the experiment from bottom to top, takes care of uncertainties and dependencies, and answers specific questions of interest. We follow a hierarchical Bayesian modelling framework and resort to MCMC for making inference. Starting out with the (unknown) number of mRNA molecules of a gene i in the target solution of a specific tissue, we follow these molecules through the process of reverse transcription and dying, mixing and centrifugation, bathing, hybridization, and washing. Conditioned on the initial amount of mRNA in the target solution, we build a hierarchical model for the hybridized target material on each spot. The image analysis step is then modelled, producing colour intensities for spots and background in two channels via (scanner and imaging specificsoftware). Finally we combine different arrays in order to answer interesting clinical questions. We show a dye-swap example to illustrate our approach and the obtainable results.

This research is supported by the Norwegian Research Council, the Cancer Society of Norway, the Norwegian Microarray Consortium and a European Union TMR programme.

Jelle Goeman, Hans van Houwelingen, Sara van de Geer

Leiden University

A global score test for differential expression of groups of genes in high-dimensional microarray data

This paper presents a global test to be used for the analysis of high-dimensional microarray data. Using this test it can be determined whether the global expression pattern of a group of genes is significantly related to some clinical outcome of interest. Such groups of genes may be any size from one single gene to all genes on the chip (e.g. known pathways, specific areas of the genome or clusters from a cluster analysis). Test results for groups of genes of different size are fully comparable, because the test gives one p-value for the group, not a p-value for each gene. Furthermore, because a group of genes is tested with a single test, multiple testing problems do not occur. The main application of the test is to investigate hypotheses about active pathways, formulated on the basis of theory or on past research. The test can also be used on all genes on the chip, e.g. for purposes of quality control.

The test is based on a goodness-of-fit score test for generalized linear models [1,2] and can handle both discreet and continuous clinical outcomes. Because of the properties of the score test, the test has optimal power against alternatives where many genes in the group have some influence on the clinical outcome. The empirical Bayesian model underlying the test is closely related to the model of penalized (linear or logistic) regression, which has been applied to microarray data with promising results [3,4]. The test result can therefore also be used as a quality label for the outcome of such an analysis on a microarray data set.

A few applications are presented in which special attention is given to visualizations of the test result that might be valuable to the biologist. These graphs are based on intuitively appealing mathematical properties of the test statistic. They can be used to visualize outlying arrays and to assess the influence of individual genes on the outcome of the test for the whole group.

References: 1: Le Cessie, S. & Van Houwelingen, H. C. (1995) Testing the fit of regression models via score tests in random effects models. Biometrics, 51, 600-614 2: Houwing-Duistermaat, J. J., Derkx, B. H. F., Rosendaal, F. R. & Van Houwelingen, H. C. (1995) Testing familial aggregation. Biometrics, 51, 1292-1301 3: Le Cessie, S. & Van Houwelingen, H. C. Ridge estimators in logistic regression. (1992) Applied Statistics, 41, 191-201 4: Eilers, P. H. C., Boer, J. M., Van Ommen, G. J. B. & Van Houwelingen, H. C. Classification of microarray data with penalized logistic regression. (2001) Proceedings of SPIE volume 4266: progress in biomedical optics and imaging, 2, 187-198

Anja von Heydebreck[1] **and Wolfgang Huber**[2]

1) Max-Planck-Institute for Molecular Genetics, Dept. of Computational Molecular Biology, Berlin, Germany, E-mail: heydebre@molgen.mpg.de

2) German Cancer Research Center, Department of Molecular Genome Analysis, Heidelberg, Germany. E-mail: w.huber@dkfz.de

Error modeling, data transformation and robust calibration for microarray data

Abstract: Two important topics in the analysis of microarray data are the calibration (normalization) of data from different experiments and the problem of variance inhomogeneity, in the sense that the variance of the measured intensities depends on their expectation value. A family of transformations for microarray intensity data has been proposed that makes the variance of transformed intensities roughly independent of their expectation value [1, 2]. In our approach, we incorporate our assumptions about the data into a statistical model and use a robust variant of maximum-likelihood estimation to estimate the parameters for the calibration and the variance stabilizing transformation simultaneously. In the following, we investigate the validity of the variance stabilizing transformation and the performance of the parameter estimation using simulated data.

Model, data transformation and parameter estimation

In this section, we give a brief summary of a family of transformations for microarray intensity data, aiming at calibration and variance stabilization, which is described in more detail in [2, 3].

We model the dependence of intensity measurements obtained from a microarray experiment on the true expression levels with an additive and a multiplicative error term:

$$Y_k = m_k e^{\eta} + \nu, \tag{2.1}$$

Here, m_k denotes the true transcript abundance of gene k (in arbitrary units); Y_k denotes the measured intensity, and η and ν are error terms following distributions $\mathcal{L}_\eta$ and $\mathcal{L}_\nu$ with mean zero, respectively. This error model was introduced by Rocke and Durbin [4], assuming normally distributed errors. Furthermore we assume that data from different arrays or from the different color channels of a single array can be brought onto a common scale via an affine–linear transformation that corrects for systematic differences in experiment conditions. Thus we obtain the following model for the measured intensities Y_{ki}, where k denotes the genes/probes and i denotes the samples:

$$\frac{Y_{ki} - a_i}{\beta_i} = m_{ki} e^{\eta_{ki}} + \nu_{ki}, \qquad \eta_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\eta,\ \nu_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\nu. \tag{2.2}$$

Here, a_i and β_i denote the parameters defining the affine–linear calibration transformation for sample i. For this model, the dependence of the variance of the measured intensities on their expectation values is given as follows:

$$\mathrm{Var}(Y_{ki}) = c^2 \left(\mathrm{E}(Y_{ki}) - a_i\right)^2 + \beta_i^2 \sigma_\nu^2, \tag{2.3}$$

where $c^2 = \mathrm{Var}(e^\eta)/\mathrm{E}^2(e^\eta)$ is a parameter of the distribution of $\eta \sim \mathcal{L}_\eta$. In the log-normal case, $c^2 = e^{\sigma_\eta^2} - 1$.

For a family of random variables Y_u with expectation values $\mathrm{E}(Y_u) = u$ and variances $\mathrm{Var}(Y_u) = v(u)$, an approximate variance-stabilizing transformation is obtained as follows [6]:

$$h(y) = \int^y 1/\sqrt{v(u)}\, du. \tag{2.4}$$

For a variance-mean dependence as in Eqn. (2.3), this leads to

$$h_i(y_{ki}) = \mathrm{arsinh}\, \frac{y_{ki} - a_i}{b_i}, \tag{2.5}$$

with $b_i = \beta_i \sigma_\nu / c$.

In order to estimate the parameters a_i and b_i from given data (y_{ki}), we postulate the following model on the transformed scale:

$$\mathrm{arsinh}\, \frac{Y_{ki} - a_i}{b_i} = \mu_{ki} + \varepsilon_{ki}, \qquad \varepsilon_{ki} \overset{\text{iid}}{\sim} \mathcal{L}_\varepsilon, \tag{2.6}$$

where $\mathcal{L}_\varepsilon$ has mean zero and variance c^2. Here we assume that the majority of genes have unchanged expression levels across the samples, that is, $\mu_{ki} = \mu_k$ for all samples i. Furthermore, we make the assumption that $\mathcal{L}_\varepsilon$ is unimodal and roughly symmetric (we have often observed heavier tails than those of a normal distribution in microarray data, also in the absence of differentially expressed genes). The parameters a_i and b_i are estimated with a robust variant of maximum likelihood estimation similar to *least trimmed sum of squares regression*. Loosely spoken, we alternate between maximum likelihood estimation of the parameters using only a subset of probes with little evidence for differential expression, and an updated computation of this subset as a certain proportion q_{lts} (trimming quantile) of probes with the smallest residuals for the current parameter values. Typically, $0.5 \le q_{lts} < 1$. A detailed description of the parameter estimation can be found in [3].

Properties of the variance stabilizing transformation

The derivation of the variance stabilizing transformation (2.5) involves a first–order approximation ("delta method") [3, 6]. Fig. 1 investigates how well this approximation holds for a family Y_m of random variables distributed according to the model

$$Y_m = m e^\eta + \nu, \qquad \eta \sim N(0, \sigma_\eta^2),\ \nu \sim N(0, \sigma_\nu^2). \tag{2.1}$$

Without loss of generality, $\sigma_\nu = 1$. The variance stabilizing transformation (2.5) has the form $h(y) = \text{arsinh}(cy)$ with $c^2 = e^{\sigma_\eta^2} - 1$. For large m, $Y_m \approx me^\eta$ and $h(Y_m) \approx \log(Y_m)$. The asymptotic standard deviation of $h(Y_m)$ for $m \to \infty$ is thus σ_η. In Fig. 1, the behavior of $\text{Sd}(h(Y_m))$ for finite values of m is compared against the asymptotic value for different choices of the parameter σ_η. For each plot, the function was numerically evaluated at 60 values of m through Monte Carlo integration with 10^6 samples. Even in the case of $\sigma_\eta = 0.4$, which corresponds to a probability of 5% of observing a relative error larger than $e^{2 \cdot 0.4} \approx 2.2$, the standard deviation $\text{Sd}(h(Y_m))$ does not depart from the asymptotic value by more than a factor of 1.035.

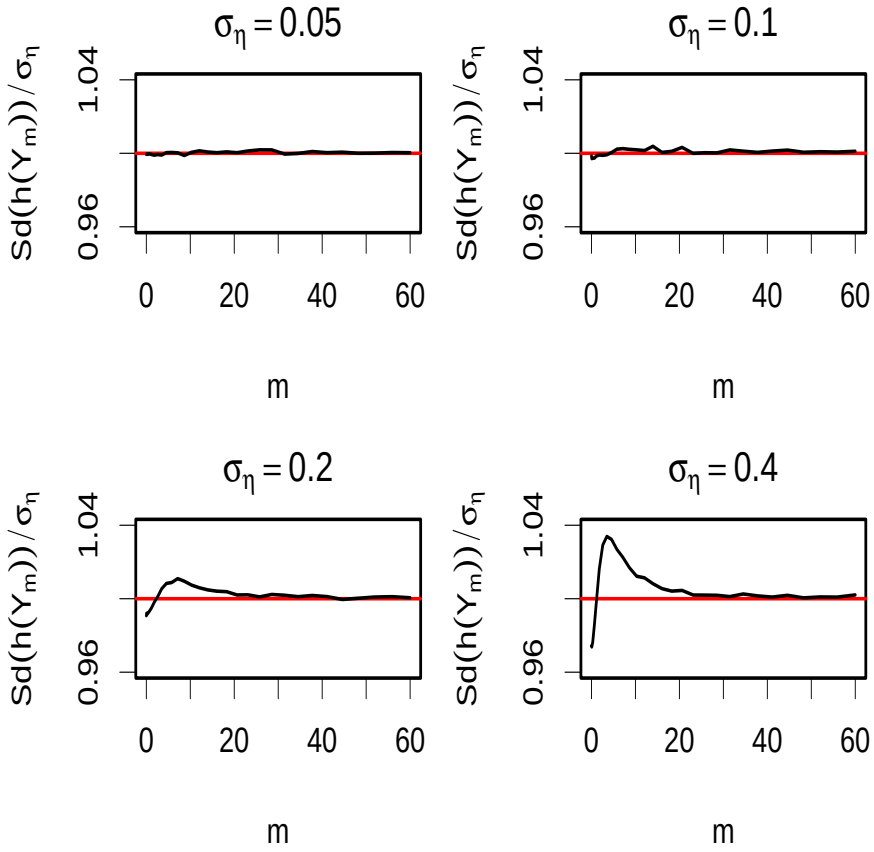

Figure 1: Validation of the approximation used in the derivation of variance stabilizing transformations. For a family of random variables distributed according to the microarray error model (2.1), the plots show the standard deviation of the transformed values $h(Y_m) = \text{arsinh}(cY_m/\sigma_\nu)$, divided by the asymptotic value σ_η, which is obtained for $m \to \infty$.

Simulation of data

To investigate the behavior of our robust variant of maximum likelihood parameter estimation [3], we ran simulation studies. Simulated data y_{ki} were generated according to model (2.6) with $\mathcal{L}_\varepsilon = N(0, c^2)$.

Values for the parameters μ_{ki} were generated as follows. First, following [5], for each gene k a value μ_k was drawn according to

$$\mu_k = \operatorname{arsinh}(m_k), \qquad 1/m_k \sim \Gamma(1,1). \tag{2.1}$$

To model the mixture of non-differentially and differentially expressed genes, indicators $p_k \in \{0,1\}$ were generated with $P[p_k = 1] = p_{\text{diff}}$. For each gene with $p_k = 1$ and for each sample $i \geq 2$ a factor $s_{ki} \in \{-1, 1\}$ was drawn with $P[s_{ki} = 1] = p_{\text{up}}$ and an amplitude z_{ki} was drawn from the uniform distribution $U(0, z_{\max})$. These were combined to obtain

$$\begin{aligned} \mu_{k1} &= \mu_k \\ \mu_{ki} &= \mu_k + p_k s_{ki} z_{ki}, \qquad i \geq 2. \end{aligned} \tag{2.2}$$

Values for the calibration parameters a_i and b_i were generated through

$$\begin{aligned} a_1 &= 0, & a_2, \ldots, a_d &\overset{\text{iid}}{\sim} U(-\Delta a, \Delta a) \\ b_1 &= 1, & b_2, \ldots, b_d &\overset{\text{iid}}{\sim} LN(0,1), \end{aligned}$$

where $\Delta a = 0.95$ roughly corresponds to the mode of the distribution of the m_k, $U(-\Delta a, \Delta a)$ is the uniform distribution on the interval $[-\Delta a, \Delta a]$, and $LN(0,1)$ is the log-normal distribution corresponding to the standard normal distribution. This yields simulated data $y_{ki} = a_i + b_i \sinh(\mu_{ki} + \epsilon_{ki})$.

The matrix (y_{ki}) was presented to the software implementation in the Bioconductor package `vsn` version 1.0 (`http://www.bioconductor.org`). The function returns the estimated transformations $\hat{h}_1, \ldots, \hat{h}_d$, parameterized by $\hat{a}_1, \ldots, \hat{a}_d$ and $\hat{b}_1, \ldots, \hat{b}_d$ (see Eqn. (2.5)), as well as the matrix of transformed data $\hat{h}_{ki} = \operatorname{arsinh}(\frac{y_{ki} - \hat{a}_i}{\hat{b}_i})$. Generalized log-ratios were calculated as

$$\Delta\hat{h}_{ki} = \hat{h}_{ki} - \hat{h}_{k1} \tag{2.3}$$

and compared to the true values

$$\Delta h_{ki} = h_{ki} - h_{k1}, \tag{2.4}$$

with $h_{ki} = \mu_{ki} + \epsilon_{ki}$, by means of the root mean squared deviation

$$\delta = \sqrt{\frac{1}{N} \sum_{i=2}^{d} \sum_{k \in \kappa} (\Delta\hat{h}_{ki} - \Delta h_{ki})^2}, \tag{2.5}$$

where κ is the set of k for which $p_k = 0$, and $N = |\kappa|(d-1)$ is the number of summands.

simulation		A	B	C	D
number of probes	n	$384, \ldots, 69120$	9216	9216	9216
number of arrays	d	2	$2, \ldots, 64$	2	2
proportion of differentially expressed genes	p_{diff}	0	0	$0, \ldots, 0.6$	0.2
proportion of up-regulated genes	p_{up}	-	-	$0.5, 1$	$0, \ldots, 1$
amplitude of differential expression	$z_{\max}$	-	-	2	2
trimming quantile	q_{lts}	$0.5, 0.75, 1$			

Table 1: Simulation parameters.

For a given set of simulation parameters, this procedure was repeated multiple times, resulting in a simulation distribution of the root mean squared error δ. This was used to obtain the error bars shown in Figs. 2, 3, and 4. The error bars are centered at the mean and extend by twice the standard error of the mean in each direction.

Simulation results

Four series of simulations were performed to investigate the influence of the number of probes, number of arrays, the choice of the trimming quantile q_{lts}, the proportion of differentially expressed genes, and the degree of asymmetry regarding the numbers of up– and downregulated genes. The parameter settings are summarized in Table 1.

The dependence of the estimation error δ on the number of probes n and the number of arrays d was investigated in simulation series A and B. The results are shown in Fig. 2. In the left plot, the number of probes n varies from 384 to 69120, with three different choices for the trimming quantile q_{lts}. ¿From the plot, a scaling of the root mean squared error approximately as

$$\delta \propto \frac{1}{\sqrt{n}} \tag{2.1}$$

can be observed. In the right plot, the number of arrays d varies from 2 to 64, again with three different values of q_{lts}. While δ does slightly decrease with d, the decrease is much slower than that with n, and does not show an obvious scaling such as (2.1). The difference between the two plots may be explained by the fact that the number of parameters that need to be estimated in order to determine the transformations (2.5) is $2d$. Thus, the number of data points per parameter remains constant when d is increased, but increases proportionally when n is increased.

The effect of the presence of differentially expressed genes on the estimation error δ was investigated in simulation series C and is shown in Fig. 3. With $q_{\text{lts}} = 1$, that is, without use of a robust estimation method, δ becomes large even in the presence of only few

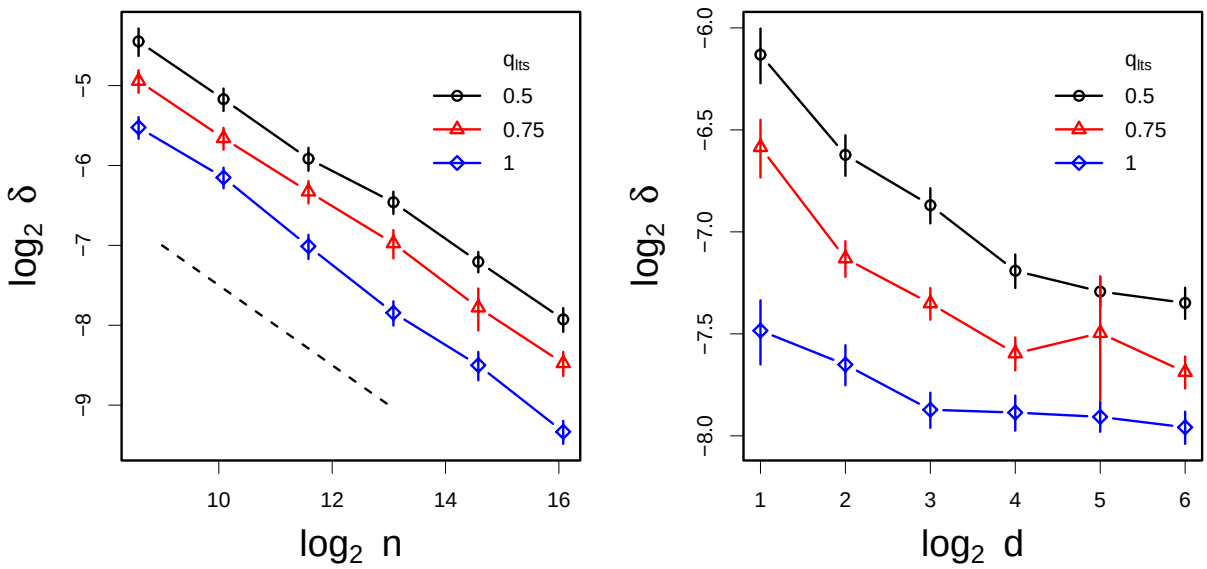

Figure 2: The root mean squared deviation δ depending on the number n of probes (left) and the number d of samples (right), for three different choices of the trimming quantile q_{lts}. The dashed line in the left plot indicates the graph of the function $f(n) = \frac{1}{\sqrt{n}}$.

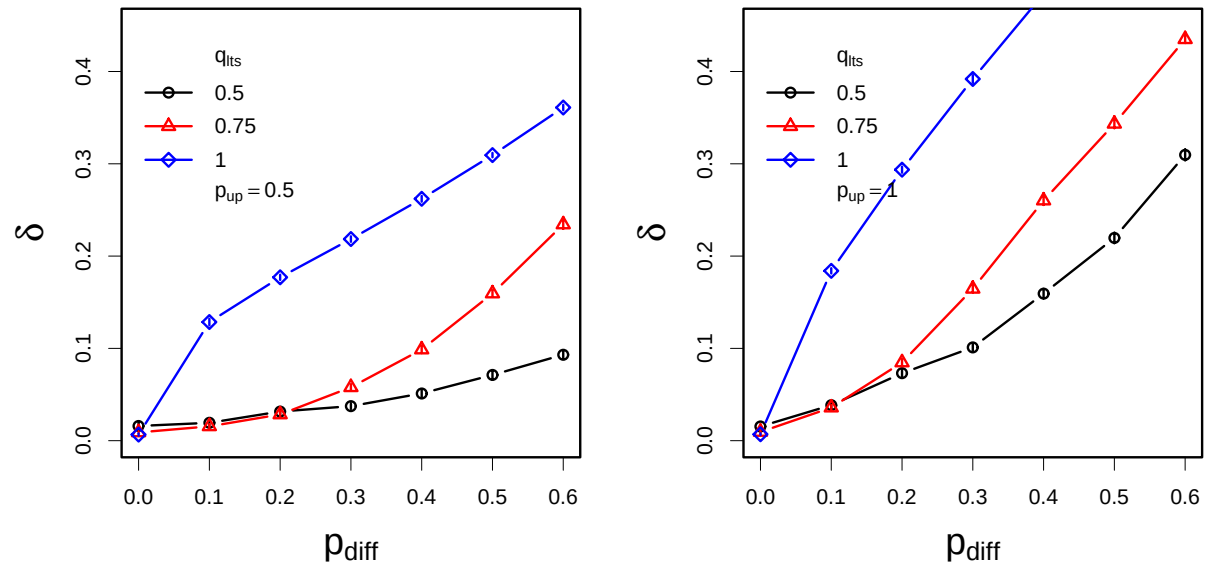

Figure 3: The root mean squared deviation δ for different proportions p_{diff} of differentialy expressed genes and three different choices of the trimming quantile q_{lts}. For the left plot, the differentially expressed genes have equal probability of being up- or downregulated; in the right plot, there are only upregulated genes.

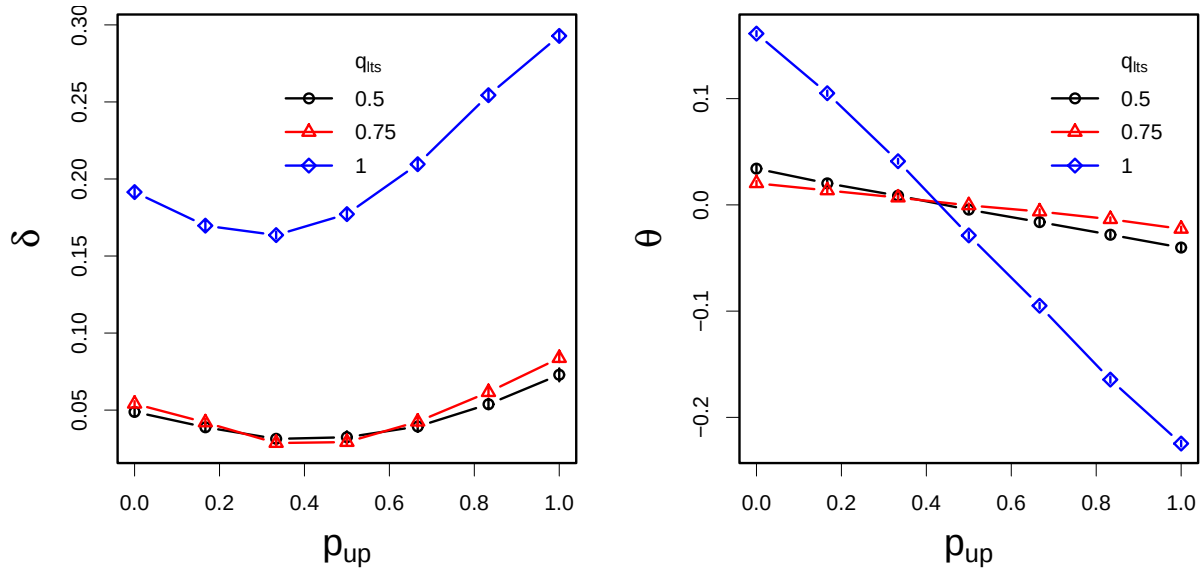

Figure 4: Estimation error δ (left) and bias θ (right) for different values of the asymmetry parameter p_{up}. Among the set of differentially expressed genes, p_{up} is the fraction of genes that are up–regulated in sample 2.

differentially expressed genes. As p_{diff} increases (starting at a value between 0.2 and 0.3), the estimation error remains smaller with trimming at the median ($q_{\text{lts}} = 0.5$) than at the 75% quantile ($q_{\text{lts}} = 0.75$). Asymmetric situations ($p_{\text{up}} = 1$, right panel) are worse than symmetric ones ($p_{\text{up}} = 0.5$, left panel), but still can be handled reasonably as long as the proportion of differentially expressed genes is not too large.

Another look at the influence of asymmetry between up- and down-regulated genes is shown in Fig. 4, which was obtained from simulation series D. Here, p_{diff} was fixed at 0.2. The right panel shows the average bias

$$\theta = \frac{1}{|\kappa|} \sum_{k \in \kappa} \left(\Delta \hat{h}_{k2} - \Delta h_{k2} \right). \tag{2.2}$$

The robust procedures ($q_{\text{lts}} = 0.5$ and 0.75) perform much better than the unrobust one ($q_{\text{lts}} = 1$). Bias θ and error δ are smallest when the fractions of up- and down-regulated genes are about the same.

References

[1] B. Durbin, J. Hardin, D. Hawkins and D. Rocke 2002. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* Vol. 18 Suppl.1, S105–S110.

[2] W. Huber, A. v. Heydebreck, H. Sültmann, A. Poustka and M. Vingron 2002. Variance Stablization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics* Vol. 18 Suppl.1, S96–S104.

[3] W. Huber, A. v. Heydebreck, H. Sültmann, A. Poustka and M. Vingron 2003. Parameter estimation for the calibration and variance stabilization of microarray data. Submitted, `http://www.dkfz.de/abt0840/whuber/`.

[4] D.M. Rocke and B. Durbin 2001. A model for measurement error for gene expression analysis. *Journal of Computational Biology* Vol. 8, 557–569.

[5] M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner and K.W. Tsui 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* Vol. 8, 37–52.

[6] R. Tibshirani 1988. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association* Vol. 83, 394–405.

Rafael A. Irizarry

Johns Hopkins University, Baltimore

Getting usable data from microarrays: The role of statisticians

In this talk I will give some examples of why I think it is important that statisticians be involved in preprocessing of microarray data. I will then describe a specific example related to preprocessing Affymetrix GeneChip high density oligonucleotide array raw data. High density oligonucleotide expression array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. Affymetrix GeneChip arrays are the most popular. In this technology each gene is typically represented by a set of 11-20 pairs of oligonucleotides separately referred to as probes. Typically 12,000 to 20,000 probe sets are arrayed on a silicon chip. RNA samples are prepared, labeled and hybridized to the arrays. Arrays are then scanned, and images produced and analyzed to obtain an intensity value for each probe. These intensities quantify the extent of the hybridization between the labeled target sample and the oligonucleotide probe. A final step to obtain expression measures is to summarize the probe intensities for a given gene in order to quantify the amount of the corresponding mRNA species in the sample. Using two extensive spike-in studies and a dilution study, we performed a careful assessment of the method of summarizing probe level data provided by the current version of the Affymetrix Microarray Suite (MAS 5.0). We found that the performance of the Affymetrix technology can be greatly improved by the use of expression measures derived from empirically motivated statistical models. The advantages of a new expression measure are assessed through bias, variance, sensitivity, and specificity. In particular, the improvements achieved by a 10-fold decrease in variability for low expression levels are demonstrated. A paper describing this example can be found on the web: `http://www.biostat.jhsph.edu/~ririzarr/papers`

Christopher Workman, Lars Juhl Jensen, Hanne Jarmer, Randy Berka, Laurent Gautier, Henrik Bjørn Nielsen, Hans-Henrik Saxild, Claus Nielsen, Søren Brunak, and Steen Knudsen

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark

A new non-linear normalization method for reducing variability in DNA microarray experiments

Background: Microarray data are subject to multiple sources of variation, of which biological sources are of interest whereas most others are only confounding. Recent work has identified systematic sources of variation that are intensity-dependent and non-linear in nature. Systematic sources of variation are not limited to the differing properties of the cyanine dyes Cy5 and Cy3 as observed in cDNA arrays, but are the general case for both oligonucleotide microarray (Affymetrix GeneChips) and cDNA microarray data. Current normalization techniques are most often linear and therefore not capable of fully correcting for these effects.

Results: We present here a simple and robust non-linear method for normalization using array signal distribution analysis and cubic splines. These methods compared favorably to normalization using robust local-linear regression (lowess). The application of these methods to oligonucleotide arrays reduced the relative error between replicates by 5-10standard global normalization method. Application to cDNA arrays showed improvements over the standard method and over Cy3-Cy5 normalization based on dye-swap replication. In addition, a set of known differentially regulated genes was ranked higher by the t-test. In either cDNA or Affymetrix technology, signal-dependent bias was more than ten times greater than the observed print-tip or spatial effects.

Conclusions: Intensity-dependent normalization is important for both high-density oligonucleotide array and cDNA array data. Both the regression and spline-based methods described here performed better than existing linear methods when assessed on the variability of replicate arrays. Dye-swap normalization was less effective at Cy3-Cy5 normalization than either regression or spline-based methods alone.

Mette Langaas

Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway, Mette.Langaas@math.ntnu.no

Estimating the number of genes truly differentiable expressed

Aim

This talk addresses two important questions in the field of DNA microarray data modelling and analysis; finding differentially expressed genes and effects, and estimating the number of true null hypotheses. Focus of the talk will be on applied aspects, and graphical illustrations will play an important role.

Finding differentially expressed genes and effects

The main objective of many DNA microarray studies is the determination of a list of differentially expressed genes. Differential expression is often sought e.g. between normal and treated cells, between different treatments, or at different timepoints during treatment. Focus may also be on effects, e.g. is there an overall difference between treatments?

Differentially expressed genes and effects can be identified using linear mixed effects models. Linear mixed effects models are models were both fixed and random effects are present. Variability both across and within genes can be taken into account. Linear mixed effects models can be used for a variety of experimental designs. Hypothesis testing can be performed by a direct generalization of simple t-tests. Linear fixed effects models were introduced in gene expression analysis by Kerr & Churchill (2000) and Kerr et al. (2002), and linear mixed effects model have been used in the analysis of gene expression by e.g. Wei et al. (2001) and Wolfinger et al. (2001).

In this presentation we look at linear mixed effects models based on transformed intensities and on transformed ratios. The models can be applied to each gene separately, to a group of genes studied together, or to all genes on the DNA microarray simultaneously. When the number of observations and effects in the linear mixed effects model is moderate, the model can be fitted using likelihood-based methods, e.g. as implemented in the NLME-library of Pinheiro and Bates (2000). With larger data sets and many random effects, Gibbs-sampling can be used.

An experiment assessing the effects of mRNA amplification on gene expression ratios in cDNA experiments, Nygaard et al. (2002), is presented in depth. The goal of this study was to evaluate if ratios were preseved by the amplification protocol under study. The question was assessed by examining data of observed intensities between two cell lines in cDNA microarray experiments using non-amplified and amplified material. Two different cell lines were considered since it is impossible to hybridize non-amplified and amplified material on the same array. We found many sources of variation present in the cDNA experiments, which were modelled as random effects in the linear mixed effects model. The parameters in the linear mixed effects model were estimated using Gibbs-sampling. From estimates of the variability in the random effects in the linear mixed effects models, a signal-to-noise evaluation criterion was constructed to arrive at a conclusion on preservation of ratios.

Estimating the number of true null hypotheses

When assessing gene significance both expression changes and precision of the expression changes are evaluated. Often several genes (e.g. all genes printed on a DNA microarray slide) are investigated at the same time, and gene significance is assessed by looking at one p-value for each gene. Genes with small p-values are declared to be differentially expressed. It is important to be able to give an overall evaluation of errors made, e.g. by giving an estimate of the number of false positive findings (genes declared to be differentially expressed which in reality are not) and the number of false negative findings (genes truly differentially expressed but not declared as such). This can be done with the aid of multiple hypothesis testing methods, and should be performed routinely when assessing differentially expressed genes in the analysis of DNA microarray data.

Popular methods focus on controlling the family-wise error rate (FWER, the probability of at least one false positive finding, i.e. the probability of at least one null hypothesis erroneously rejected) or the false discovery rate (FDR, the expected proportion of false positive findings, i.e. the expeced proportion of the rejected null hypotheses erroneously rejected) of Benjamini and Hochberg (1995). A related quantity is the false non-discovery rate (FNR) of Genovese and Wassermann (2001), which look at the expected proportion of false negatives, i.e. the expected proportion of non-rejected null hypotheses erroneously not rejected.

When estimating the FDR, an estimate of the number of true null hypotheses must be given. But, the number of true null hypotheses is also an interesting quantity in itself. In the current presentation we look at two different methods for estimating the number, m_0, of true null hypotheses when m related hypotheses are tested simultaneously. The methods are based on observed p-values, which are assumed to be independent.

The first estimator was suggested by Schweder and Spjøtvoll (1982). This estimator has been used in the estimation of the FDR by Storey (2002). The estimator is intuitively simple, but includes a meta-parameter that is difficult to estimate.

We also consider a method based on estimating the distribution of the p-values. More specifically, we find a decreasing nonparametric maximum likelihood density estimate of the distribution of p-values.

The methods are compared and depictured graphically in a simulation study. Estimation is also done for the Nygaard et al. (2002) data.

This work is joint with Egil Ferkingstad and Bo Lindqvist.

References

Benjamini & Hochberg (1995) *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing*, JRSS B, 57, 1, 289-300.

Genovese & Wasserman (2001) *False Discovery Rate*, Technical report, Carnegie Mellon University.

Kerr& Churchill *Analysis of variance for gene expression microarray data*, Journal of Computational Biology, Vol. 7, pp. 819-837.

Kerr, Afshari, Bennett, Bushel, Martinez, Walker & Churchill(2002) *Statistical analysis of a gene expression microarray experiment with replication*, Statistica Sinica.

Nygaard, Løland, Holden, Langaas, Rue, Liu, Myklebost, Fodstad, Hovig & Smith-Sørensen (2002) *Effects of mRNA amplification on gene expression ratios in cDNA experiments estimated by analysis of variance*, submitted.

Pinheiro & Bates (1990) *Mixed-Effects Models in S and S-PLUS*, Springer.

Schweder & Spjøtvoll (1982) *Plots of P-values to evaluate many tests simultaneously*, Biometrika, 69, 3, 493-502.

Storey (2002) *False discovery rates: Theory and application to DNA microarrays*, PhD thesis, Department of Statistics, Stanford University.

Wei, Riley, Wolfinger, White, Passador-Gurgel and Gibson (2001) *The contribution of sex, genotype and age to transcriptional variance in Drosphila melanogaster*, Nature Genetics, Vol. 29, pp. 389-395.

Wolfinger, Gibson, Wolfinger, Bennett, Hamadeh, Bushel, Afshari & Paules *Assing Gene Significance from cDNA Microarray Expression Data via Mixed Models*, Journal of Computational Biology, Vol. 8, pp. 625-637.

Volkmar Liebscher

Institute of Biomathematics and Biometry, GSF — National Research Centre for Environment & Health, Neuherberg, Germany

Stochastic Modelling and Quality Control for Gene Expression Data

Introduction

cDNA microarray experiments are a major tool in functional genomics. They are designed to analyse gene expression profiles of a given organism in order to understand its gene regulatory network. Despite the enormous potential of this kind of experiment, the possibility to extract useful information from the data is limited by an abundance of noise generated in the course of this complex experiment. Any improvement of the experimental design and the analysis of the data must be based on a deeper understanding of the relation of data and the abundance of gene transcripts.

Microarray Experiments

From two different types of cells, cell fluid containing mRNA is extracted. The mRNA sequences are labelled by two different markers (*green/red*) and the samples are mixed. Small amounts of this mixture are exposed to the respective spots, where probes are situated. Ideally, only specific sample sequences complementary to the probes bind to it. A subsequent washing procedure removes all unbound mRNA from the spot, the markers of the remaining ones produce a signal which can be measured to quantify the amount of specific mRNA on the respective spot and hence to screen the relative abundance of mRNA in the two samples.

In this highly involved procedure there are different sources of noise. Possibly, the images look like Fig. 5, although this example is aN extreme one. Examples of noise include biological variation (e.g. the cells used to extract mRNA are not in the same state), chemical variation (e.g. binding energies may depend on fluourescence markers) and measurement variation (e.g. machine parameters are not stable over time). The latter could be analysed in the course of repeated experiments, see e.g. [8].

Cross Hybridization

One particular source of error are so called *cross hybridizations*, i.e. the binding of unspecific mRNA to the probe. Here, unspecific means that the mRNA sequence is not

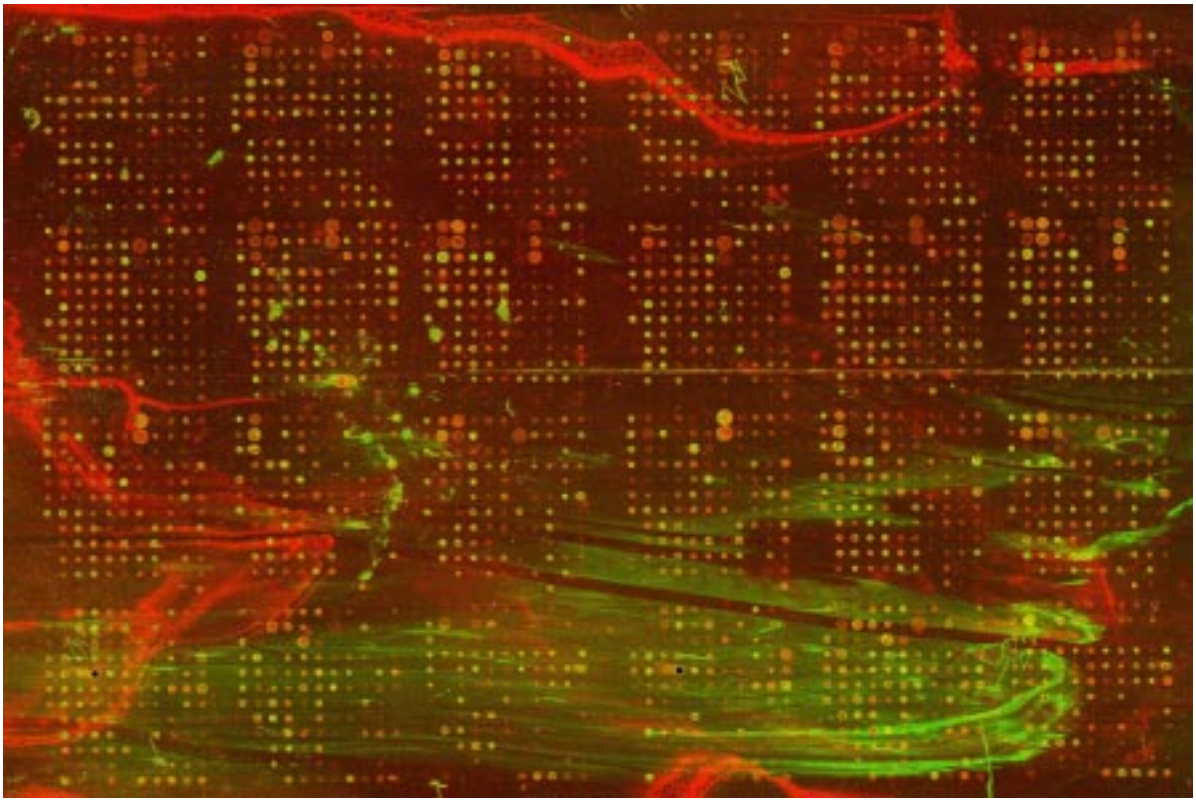

Figure 5: Red/green image from a microarray experiment (raw data).

really close to the complement of the respective probe sequence. The occurence of a relevant number of unspecific bindings can cause that the signal finally measured is not proportional to the number of specific mRNA in the sample since all hybridized mRNA sequences contribute to the signal.

Therefore there is need to validate specificity of hybridized mRNA. Clearly, this requires specially designed experiments, one possibility is described in the next section. We want to decide from the additional experimental data thus provided whether the signal corresponding to a given spot is produced exclusively by hybridization of specific mRNA and — in case it is not — which conclusions can still be drawn from the data.

Validation of Goodness of Spots — An Experiment

In this work, we are concerned with a special experimental technology suggested in [1] to assess cross hybridization.

Initially, a usual cDNA microarray experiment is performed. The intensities of the spots are measured and stored. Then, the chip is exposed to a washing procedure (using formamide) removing some of the hybridized mRNA from the chip. Again, the intensities are measured. This is repeated several times increasing the concentration of formamide

from step to step. The idea is twofold:

- The higher the binding energy of mRNA and probe the higher is the concentration of formamide necessary to remove it from the chip.
- The specific mRNA is the one with highest binding energy, i.e. it fits best to the probe.

Now we plot formamide concentration against signal intensity. For spots with no relevant expression signal we expect just noise (Fig. 6, left). If there is no cross hybridizations we expect one single decrease of intensity located at the binding energy of the specific mRNA (Fig. 6, middle). For spots with cross hybridizations we expect several decreases in intensity corresponding to different binding energies the last of which corresponds to the specific sequence (Fig. 6, right).

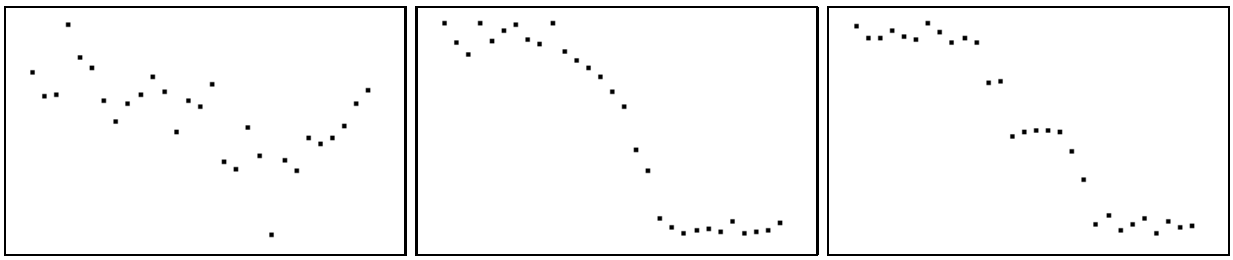

Figure 6: Experimental Data: Noise, Good Spot, Bad Spot

Modelling of the Washing Process

To get a better theoretical understanding of the washing curves we model the washing process by a yet simple urn model.

To compute the probability that a bound cDNA molecule remains bounded after washing we assume

1. There is a number of existing bonds, say M.
2. Only one bound is broken by a detergent molecule.
3. The detergent is randomly distributed, its interaction with the bonds is random.
4. There is the possibility that furthers bonds break by thermal effects.
5. The cDNA remains bounded to the array unless the number of broken bounds exceeds another number K (usually, $K \sim 1$ is very small).

This implies that the number of non-broken bonds is binomially distributed as $\mathrm{B}(M, p)$. There $p = \mathrm{e}^{-\ln\varepsilon - c_f(1-\mu)}$ where μ and ε are the probabilities that a certain detergent molecule breaks a bond and a bond breaks thermally respectively, and c_f is the concentration of detergent in the washing solution.

Now we implement the washing schedule into the same model. Assuming linear increase in concentration of formamide and the possibility that formamide molecules remain on the array (already breaking bonds) from one step to the other implies that the probability that a mRNA molecule remains attached to the array after the kth washing step is $\mathrm{B}(M, p_k)(\{0, \ldots, K\})$ where

$$p_k = \mathrm{e}^{C_1 + C_2 k + C_3 k^2}$$

for suitable constants C_1, C_2, C_3 .

Adaptation to Data

The analysed data concerned differential expression of 20 000 genes in testes from 105 day old mice vs. whole embryos at day 10.5 (1 replicate). There were recorded the data from 29 washing steps. More details of the experiment are contained in [1].

First we analyse whether the data show the behaviour suggested by our model above. To cope with the possibility of bad spots (with significant occurence of cross hybridization) and outliers in the data, we normalized the washing curves robustly requiring the order statistics to fulfil $x_{22:29} \stackrel{!}{=} 1$ and $x_{6:29} \stackrel{!}{=} 0$. Thereafter, we computed the pointwise median of the normalised washing curves over all 16 606 significantly expressed spots. This should reveal the basic form of a washing curve since the shape is stable under the median transformation and it is estimated that around 80% of the spots are good, with similar transition points. Further, exploring this median curve, it is possible to discover other systematic effects in the data, see Fig.7.

The best least squares fit of the above model to these median data is shown in Fig.7. Surely, the model has problems to adapt to the plateau on the right. Clearly, further parameters need to be introduced to consider mixtures of more than one such signals (due to cross hybridization) too. This makes it unreasonable to fit such (though validated) model to the data to detect spots with cross hybridization.

Detecting Jumps

On the other hand, both the data and our model show relatively sharp transitions in the intensities. Thus, *jumps carry the essential information* in these data.

Therefore, we use instead of a fully parametric the much simpler approach to fit piecewise constant functions to the data [7]. Doing this, we do not aim at perfect fits but on

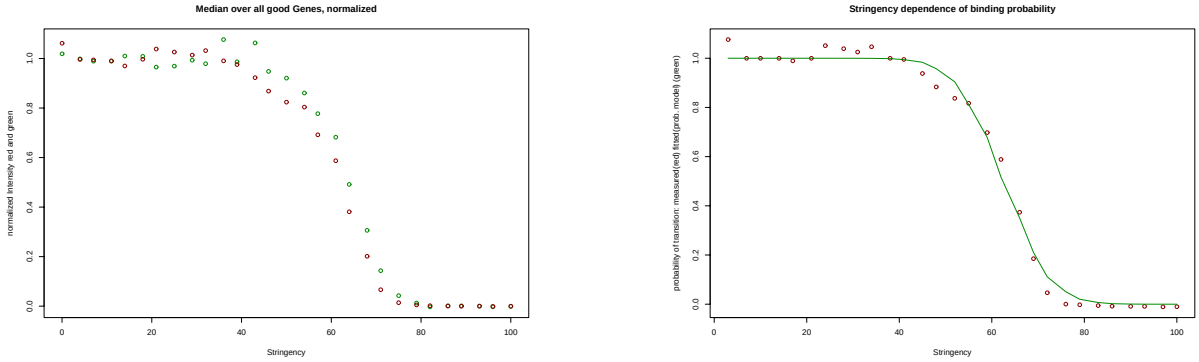

Figure 7: left: Median of normalized washing curves, red and green channel; right: Best fit of red channel to binomial model with $M = 50$, $K = 12$.

extracting *essential features* following DAVIES [2]. Basically, we choose that fit x to data y which minimizes

$$H(x, y) = \gamma\, J(x) + \sum_i \varphi(x_i - y_i)$$

where $J(x) := \#\{i : x_i \neq x_{i+1}\}$ is the number of "jumps", γ is a constant controlling "smoothness" of the fit and φ is a function controlling fidelity to the data. For the original proposal in [7], $\varphi(u) = u^2$, we obtain a kind of least squares fit. On the other side, robustness with respect to noise suggests to use more slowly increasing functions like $\varphi(u) = |u|$, $\varphi(u) = \min(u^2, 1)$, $\varphi(u) = \min(|u|, 1)$ or HUBER's proposal $\varphi(u) = \begin{cases} u^2 & |u| \leq 1 \\ 2|u| - 1 & |u| > 1 \end{cases}$, see [4, 5]. For all of these choices, the minimizers of $H(\cdot, y)$ are easily computable for all γ simultaneously, see [6].

The critical point is the choice of γ where different strategies can be adopted [6]. It is basic to this kind of model that the best fit as a function of $\gamma$for fixed data y is piecewise constant. It is reasonable to choose that reconstruction which belongs to the largest γ interval of constancy [6], see Fig. 8.

The various ways to interprete data in a piecewise constant fashion were implemented in Oberon under the software platform ANTSINFIELDS [3].

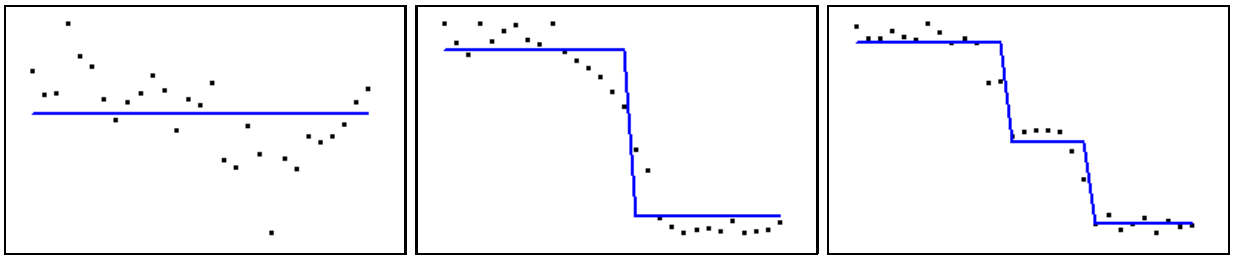

Figure 8: Potts reconstructions for the signals from Fig.6

Results and Perspectives

Although this is work in progress, we want to emphasise the following points.

- There are outliers and there is no hypothesis available about the exact distribution of noise. Thus, noise must be treated robustly.
- Usually, only a few replicates (sometimes a single one) are available, therefore we need contextual information (i.e. monotonicity, sharp transitions) to enhance inference.
- Contextual information is provided by both, a *probabilistic model* and *empirical analysis*.
- Due to the contextual information thus extracted, Potts models and/or Taut String methods are promising tools to investigate these data.
- Our goal is to quantify cross hybridizations and to use this knowledge to investigate cDNA microarray data with higher degree of reliability.

Acknowledgements

This work was jointly done with J. Beckers, St. Brandt, A.L. Drobyshev, F. Friedrich, A. Hutzenthaler, A. Kempe, A. Martin, G. Winkler and O. Wittich.

It is founded by BMBF within the "Bioinformatics for the Functional Analysis of Mammalian Genomes" (BFAM) project. The research on edgepreserving smoothing is supported also by the DFG-Schwerpunkt 1114 "Mathematical methods for time series analysis and digital image processing" and EU-RTN "Harmonic Analysis and Statistics in Signal and Image Processing" (HASSIP).

References

[1] A.L. Drobyshev, Ch. Machka, M. Horsch, M. Seltmann, V. Liebscher, M. Hrabè de Angelis, and J. Beckers. Specificity assessment from fractionation experiments (SAFE): A novel method to evaluate microarray probe specificity based on hybridization stringencies. *Nucleic Acids Res.*, **31**(2):1–10, 2003.

[2] P.L. Davies. *Data features.* Stat. Neerl., **49**(2):185–245, 1995.

[3] F. Friedrich. AntsInFields*: A Software Package for Random Field Models and Imaging.* Institute of Biomathematics and Biometry, National Research Center for Environment and Health, Neuherberg/München, Germany, 2002. URL: `http://www.AntsInFields.de`.

[4] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics*. Wiley & Sons, New York, 1986.

[5] P.J. Huber. *Robust Statistics*. Wiley & Sons, New York, 1981.

[6] A. Kempe. *Statistical analysis of the Potts model and applications in biomedical imaging*. PhD thesis, Institute of Biomathematics and Biometry, National Research Center for Environment and Health, University of Munich, Germany, 2003.

[7] V. Liebscher and G. Winkler. *A Potts model for segmentation and jump-detection.* Proceedings of the S^4G-conference, Prague, 1999.

[8] Y. Tu, G. Stolovitzky and U. Klein. Quantitative noise analysis for gene expression microarray experiments. PNAS **99**(22):14031–14036, 2002

[9] G. Winkler and V. Liebscher. Smoothers for discontinuous signals. *J. Nonpar. Statist.*, 14(1–2):203–222, 2002.

Claus-Dieter Mayer

BioSS office Rowett Research Institute, Aberdeen, Scotland

Least trimmed squares methods for microarray normalization

Following the initial image analysis normalisation and standardisation is the second crucial preprocessing step in the analysis of results from a microarray experiment. Normalisation is necessary because the intensities measured on microarrays are not equal to the original mRNA abundance but are disturbed by a series of systematic experimental effects. These include effects depending on the array, dye, spot location, print-tip, intensity level, scanner settings and more. A good overview over different normalisation strategies is given in Yang et. al. [9]. The aim of these methods is to eliminate the different effects and to transform (differential) expression intensities from different arrays to quantities that are comparable. In a first step all intensities are re-scaled by a common transformation. The standard scale used in most analyses is the log-scale, one of the reasons for this being its variance stabilising behaviour, i.e. a linear trend in the signal intensity vs. standard deviation relationship is removed. Recent articles have suggested that in some experiments other transformations yield better variance stabilisation (cf. Huber et al. [5] or Durbin et al. [2]).

This present lecture assumes that an appropriate scale for the intensities has been chosen (in most applications this will be the log-scale) and studies the subsequent calibration that is needed to compare intensities measured on different channels and arrays. Methods that have been proposed for this step can be roughly divided into 3 categories:

1. **Use of control spots:** One strategy to deal with the normalisation problem is to include spots on the array which control for the different experimental effects. A typical example for this are so-called *housekeeping genes*. These are positive controls, i.e. genes which are known to be highly expressed on a constant level under different experimental conditions. Negative controls, i.e. genes which are not expected to be expressed at all, are also commonly used. Spiking is another control technique. Here spots with synthetic cDNA are included on the array and the samples are spiked with a known amount of the corresponding DNA, which should give a constant spot-intensity. These and other control strategies obviously depend very much on the specific experiment and will not be discussed in detail here. A general problem is, that the controls often do not behave as expected and show great variability. Although control spots can be a very helpful tool to monitor the experiment and its analysis, a calibration based only on these spots is highly questionable.

2. **Global mean normalisation (and related methods)**: This method assumes that the total mRNA abundance in all samples should be the same and for this reason divides the spot intensities by the mean (or sum) of all intensities measured on the same channel. The resulting normalised intensities should then be proportional

to the percentage of the total abundance expressed by the corresponding genes and thus be comparable across channel and arrays.

This procedure assumes a multiplicative effect on the original scale which turns into an additive effect, when the we consider the log-intensities which are usually studied. An obvious variation is to subtract the mean on the log-scale, which corresponds to a division by the geometric mean on the original scale. Often the mean is also replaced by a more robust measure of location as for example the median (or some other empirical quantile).

ANOVA-type methods as discussed by Kerr et al. [6] can be viewed as a generalisation of global mean normalisation methods. Here the additive array and channel effects are estimated simultaneously with the gene effects of interest.

3. **Regression based methods:** One obvious disadvantage of global mean normalisation is that it heavily relies on the assumption of additive effects. In many experiments there is a strong indication that this assumption is violated. For example dye effects often seem to be intensity dependent. For this reason methods have been studied that fit a regression curve to the scatterplot of the two channel intensities for cDNA-arrays. One problem here is that standard regression has to make an arbitrary choice which of the two channels should be treated as the explanatory variable and which one is the response variable. Another problem is that ideally the curve should be only fitted to the non-differentially expressed genes, but one doesn't know which ones these are. For this reason methods should be robust against presence of a certain amount of differential genes.

   Sapir and Churchill [10] tackled both of these issues simultaneously by using robust orthogonal linear regression. A different approach was chosen by Terry Speeds group, cf. [9], who prefer to rotate the channel 1 vs. channel 2 scatterplot by 45 degrees and to study a plot of the mean of the log-intensities vs. the log-ratio ("MA plot"). Any regression in this setting automatically treats both channels symmetrically. These authors fit a loess curve to this plot, which is a non-parametric (and thus non-linear) local regression method, that also has good robustness properties. This approach is used very frequently nowadays.

   For Affymetrix chips and other one-channel experiments there is no obvious counterpart of this loess normalisation. Bolstad et al. [1] discuss several methods to normalise these data, one of which is *cyclic loess*, where a loess regression for a MA-Plot is iteratively applied to all pairwise combinations of arrays.

   An alternative method for one-channel data has been studied by Rattray et al. [4] who assume a linear relationship between the unknown true gene expression levels and the ones measured on the arrays. Assuming Gaussian errors the parameters of this models are then estimated by maximum likelihood. This approach is also known as a *total least squares* (TLS) method and is directly related to a principal component analysis and the main calibration parameter is the first *eigenarray*. If only two arrays are present this method is identical to the orthogonal linear regression approach of Sapir and Churchill apart from the robustification.

This presentation will built up on the aforementioned methods. As already indicated all methods mentioned in 2 and 3 either assume that only very few genes will differ in expression between the different channels/arrays or they need robustness properties to eliminate the effect of differential genes. We think that there are two quite different reasons why robustness is needed in microarray statistics. The first reason is that due to the complicated process that generates these data they are prone to contain *outliers*, by which we mean unrepresentative values that might be caused by experimental failure, dirt on the array, scanning mistakes etc. ....

As (hopefully) these outliers only make up a minor part of the data any statistical procedure with a reasonable *breakdown point* (that is the amount of corrupted data, that are necessary to cause a complete collapse of the procedure) should be able to cope with this problem. In global mean normalisation for example replacing the mean (breakdown point of 0, i.e. one extreme observation can cause a collapse) by the median (breakdown point of 50 %) should ensure that the normalisation is not seriously affected by such outliers.

The other reason why robustness is needed is the before-mentioned fact that microarray data usually consist of a mixture of genes which do not change in expression under the experimental conditions of interest and others that do (these are typically the genes one is interested in). These differential genes are not outliers in the sense of the previous paragraph and their presence demands a procedure that is not only robust but also able to estimate parameters of a subpopulation of the complete data. The method of *least trimmed squares* (LTS) is specifically tailored for a situation like this. LTS is modification of a least squares fit of a model, where now only a certain percentage of the data with the smallest residuals is used for the fit. This approach has first been introduced by Rousseeuw in 1984 and is discussed in detail in Rousseeuw and Leroy [7]. In microarray analysis LTS methods can be applied whenever one can give an upper bound for the percentage of expected differential genes.

In the case of global mean normalisation this approach yields an alternative location estimator. We will discuss this simple situation to gain some insight in the principle method. We would like to stress that in our view this LTS location estimator is not an alternative method to estimate the mean of a distribution but is estimating a functional, that in general is different from the mean. For mixture distributions this functional for example can be the mean of the component with the smaller variance.

Least trimmed squares methods have already been used for microarray normalisation in Huber et al. [5], who estimate the parameters of their arsinh-transformation this way or in Faller et al. [3]), who similarly estimate parameters in their normalisation procedure. We will discuss several other potential application of LTS. In particular we will show how it can be used for normalisation of a series of one-channel experiments. Asymptotic properties of these methods will be discussed as well as algorithmical aspects, see also Rousseeuw and Van Driessen [8].

References

[1] B.M. Bolstad, R. A. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *submitted to Bioinformatics*, 2002.

[2] B.P. Durbin, J.S. Hardin, D.M. Hawkins, and D.M. Rocke. A variance-stabilizing transormation for gene-expression microarray data. *Bioinformatics*, 18:105–110, 2002.

[3] D. Faller, J. Timmer, H.U. Voss, Honerkamp and U. Hobohm "Optimal" normalization of DNA-microarray data *Talk at Workshop: Biometrical Analysis of Molecular Markers, Heidelberg 2001*, `http://webber.physik.uni-freiburg.de/~fallerd/Talks/heidelb.pdf`

[4] M. Rattray, RN. Morrison, D.C. Hoyle and A. Brass. DNA microarray normalisation, PCA and a related latent variable. *Technical Report*, 18:576–584, `http://www.cs.man.ac.uk/~magnus/magnus.html`, 2001.

[5] W. Huber, A. von Heydebreck, H. Sültmann, A. Proustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to quantification of differential expression. *Bioinformatics*, 18:96–104, 2002.

[6] M.K. Kerr, M. Martin, and G.A. Churchill. Analysis of variance for gene expression microarray data. *J Comput Biology*, 7:819–837, 2000.

[7] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection.* Wiley & Sons, New York, 1987.

[8] P.J. Rousseeuw and K. Van Driessen. Computing LTS Regression for Large Data Sets. *Technical Report*, University of Antwerp, 1999, ftp://win-ftp.uia.ac.be/pub/preprints/99/Comlts99.pdf

[9] Y.H. Yang, S. Dudoit, P. Luu, and T.P. Speed. Normalization for cDNA microarray data. In M.L. Bittner, Y. Chen, A.N. Dorsel, and E.R. Dougherty, editors, *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE*, 2001.

[10] M. Sapir and G.A. Churchill. Estimating the posterior probability of differential gene expression from microarray data. *Poster*, 2000, `http://www.jax.org/research/churchill/pubs/marina.pdf`.

Yudi Pawitan[1], Judith Björhle[3], Sara Wedren[1], Keith Humphreys[1], Lambert Skoog[3], Fei Huang[2], Lukas Amler[2], Peter Shaw[2], Per Hall[1], Jonas Bergh[3]

[1]Department of Medical Epidemiology, Karolinska Institutet, Stockholm; [2]Bristol Myers Squibb, New Jersey, USA; [3]Karolinska Hospital, Stockholm

Survival analysis using gene expression data

We consider the association between gene expression and survival in a cohort of breast cancer patients, who were operated in Karolinska Hospital between 1994 and 1996. We have so far obtained gene expression profiles for 134 patients using Affymetrix genechip U133A and B. These patients were followed for an average of 5 years, during which 38 of them had relapse or died. We analyse 11,219 genes that were expressed in more than half of the patients. To identify genes that are individually associated with survival we compute the logrank test (the score test from the Cox partial likelihood) for each gene separately. However, rather than using the standard P-value, the significance of the test is evaluated using an empirical Bayes methodology adapted to survival analysis setting. In this methodology, which takes the multiplicity of tests into account, the statistical significance is evaluated in terms of a posterior probability of a true association given the observed test statistic and this probability can be interpreted as conditional true discovery rate.

Methodology

Our main objective is to study the association between gene expression and survival, which will be done on gene-by-gene basis. Multiplicity issue creates a problem in assessing the significance of observed differences in gene expression. In particular the standard P-value is not meaningul, and its simple minded adjustment such as using Bonferroni method is too conservative. We will follow Efron et al.'s (2001) by supposing that there is an unknown proportion p_0 of genes that are not associated with survival, and a proportion $p_1 = 1 - p_0$ that are. Denote by A the indicator whether or not a gene is associated with survival. Let Z be the logrank statistic and assume that the conditional distribution of Z given $A = 0$ is $f_0(z)$, and of Z given $A = 1$ is $f_1(z)$, with unknown $f_0(z)$ and $f_1(z)$. The observed collection of Z's is a sample from

$$f(z) = p_0 f_0(z) + p_1 f_1(z).$$

What we want is $P(A = 1|z)$, which can be interpreted the true discovery rate. Using Bayes formula

$$P(A = 1|z) \quad = \quad 1 - P(A = 0|z)$$

$$
\begin{aligned}
&= 1 - \frac{P(A=0)f(z|A=0)}{f(z)} \\
&= 1 - p_0 \frac{f_0(z)}{f(z)}.
\end{aligned}
$$

The density $f(z)$ can be estimated from the data, while the 'null' density $f_0(z)$ can be generated by a permutation argument. The permutation step is a key step that will vary from application to application. Assuming we can estimate f_0, the unknown p_0 can be replaced by an upperbound:

$$
\begin{aligned}
P(A=0|z) &= p_0 \frac{f_0(z)}{f(z)} \leq 1 \\
p_0 &\leq \frac{f(z)}{f_0(z)}, \quad \text{for all } z \\
&\leq \min_z \frac{f(z)}{f_0(z)}
\end{aligned}
$$

In our application we have censored survival data (y_i, x_i, δ_i), where y_i is the time to event or last followup of the ith subject, x_i is the gene expression value and δ_i the event indicator. As usual $\delta_i = 1$ if y_i is an event time and zero otherwise. For general independent censoring mechanism a simple minded permutation of x_i does not work, since censoring might depend on x_i. Intead we generate $f_0(z)$ by the following scheme:

- at kth event time we setup

$$
\begin{aligned}
\text{Risk set} &= \{i_1, \ldots, i_{n_k}\} \\
\text{Event set} &= \{\delta_{i_1}, \ldots, \delta_{i_{n_k}}\} \\
\text{Covariate set} &= \{x_{i_1}, \ldots, x_{i_{n_k}}\}
\end{aligned}
$$

- randomly permute the event set and match it to the covariate set.
- repeat the procedure at all event times, and combine all the sampled event set to make up new (permuted) dataset
- compute the logrank test for all genes.

Each round of this procedure generates a full list of permuted logrank tests $\{Z_1^*, \ldots, Z_g^*\}$, where g is the total number of genes. The whole procedure is repeated a number of times and $f_0(z)$ is then estimated from the whole collection.

Results

We apply the methodology to a dataset obtained from a cohort of breast cancer patients who were operated in Karolinska Hospital between 1994 and 1996. We obtained gene

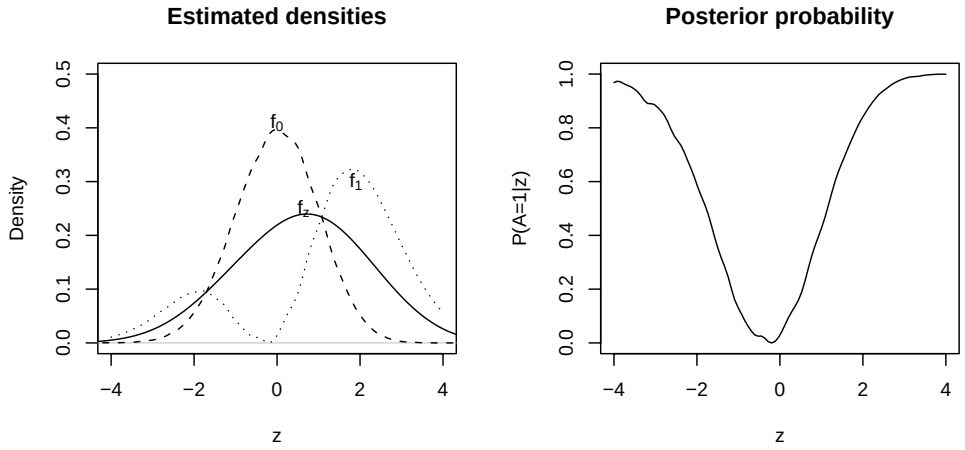

Figure 9: *(a) Estimated densities of* $f_0(z)$*,* $f_1(z)$ *and* $f(z)$*. (b) Estimated posterior probability* $P(A = 1|z)$*.*

expression profiles for 134 patients using Affymetrix genechip U133A. These patients were followed for an average of 5 years, during which 38 of them had relapse or died. We analyse 11,219 genes that were expressed in more than half of the patients. Figure 1 shows the nonparametric estimate of $f_0(z)$, $f_1(z)$ and $f(z)$ and the posterior probability $P(A|z)$. The probability immediately provides a statistical assessment of observed differences. For example, $P(A = 1|z = 3.6) = 1$ meaning that if a gene has an observed $Z = 3.6$ we are certain it is associated with survival. The probability is greater than 95% if $z < -3.6$ or $z > 2.7$.

References

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96, 1151–1160.

Mark Reimers

Karolinska Institutet, Stockholm

Issues in probe level analysis of affymetrix data

We will discuss issues in normalization and model building from the multiple-probe Affymetrix chips. We investigate whether additional information such as AT content influences normalization.At present the best multi-chip models, dChip and RMA ignore MM values. We find that MM and PM values are highly correlated for many probes, and investigate whether MM information can be profitably used in a robust multi-chip model.

Mats Rudemo

Department of Mathematical Statistics, Chalmers University of Technology, SE – 412 96 Göteborg

Empirical Bayes analysis of variance models for microarray data

Variance estimation for estimated gene effects is crucial for identifying differentially expressed genes in microarray experiments. Current methods, cf. [1]–[3], are briefly reviewed and a new variance component model is suggested in [4] for cDNA microarray trials. One major component in this model corresponds closely to a binomial model for the competition between the two types of cDNA targets labelled with different fluorescent dyes. For optimal weighting of global variance estimates and individual gene-based variance estimates an empirical Bayes procedure based on cross-validation is suggested. The suggested model is applied to a dye-swap experiment and it is compared with a number of other methods in a simulation study with varying numbers of slides and varying gene-specific variances. One conclusion is that the suggested method is roughly equivalent to a straightforward t-test with twice as many slides.

Models for microarray data with two treatments

Suppose that we have microarray data from S slides, denoted $s = 1, \ldots, S$. For each slide there are two treatments $t = 1, 2$. Let

$$Z_{gts}, \quad g = 1, \ldots, G, \; t = 1, 2, \; s = 1, \ldots, S, \tag{2.1}$$

denote the observed intensity value for the spot corresponding to gene g and treatment t in slide s, where G is the number of genes. We assume here that each spot corresponds to one gene. Let Y_{gs} denote the observed relative effect of treatment 1 compared to treatment 2 on a log-scale

$$Y_{gs} = \log \frac{Z_{g1s}}{Z_{g2s}}, \tag{2.2}$$

where log denotes natural logarithms, and let

$$x_{gs} = -\log(\frac{1}{2}(\frac{1}{Z_{g1s}} + \frac{1}{Z_{g2s}})) \tag{2.3}$$

be the log-harmonic mean intensity. The variance structure can be modeled in several ways. Let σ^2_{gs} denote the variance of Y_{gs} after normalization. As variance model we suggest

$$\sigma^2_{gs} = \sigma^2 \exp(-\alpha_1 x_{gs}) + \alpha_2. \tag{2.4}$$

A binomial model

The model (2.4) for the variance contains as a special cases a simple binomial model. Disregard here normalization and suppose that there in a spot are N fluorescently tagged cDNA targets that will give a signal contribution and that N_1 of them give a contribution to the treatment signal $Z_1 \approx cN_1$ and $N_2 = N - N_1$ of them give a contribution to $Z_2 \approx cN_2$. Due to competition among the two types of cDNA targets labelled with

different fluorescent dyes it is natural to condition upon $N = n$ and assume that N_1 is Binomial(n, p) given $N = n$ with a proportion p corresponding to the gene tested at the regarded spot. Then

$$\begin{aligned} \mathrm{Var}(Y_{gs}) &\approx \mathrm{Var}(\log \frac{N_1}{n - N_1}) \approx \mathrm{Var}(N_1) \left(\frac{\partial}{\partial x} \log \frac{x}{n-x} \Big|_{x=np} \right)^2 \\ &= \frac{1}{np} + \frac{1}{n(1-p)} \approx \frac{1}{N_1} + \frac{1}{N_2} \approx c \left(\frac{1}{Z_{g1s}} + \frac{1}{Z_{g2s}} \right) \\ &= 2c \exp(-x_{gs}), \end{aligned} \tag{2.5}$$

which corresponds to (2.4) with $\alpha_1 = 1$ and $\sigma^2 = 2c$. In examples with real data it turns out that a weighted average of (2.4) and a gene-specific variance component gives a good description of the observed variance estimates. The optimal weighting is estimated by an empirical Bayes procedure.

References

[1] Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

[2] Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of American Statistical Association* **96**, 1151–1160.

[3] Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

[4] Rudemo, M., Lobovkina, T., Mostad, P., Scheidl, S.J., Nilsson, S., and Lindahl, P. (2002). Variance models for microarray data. Report 2002:6, Mathematical Statistics, Chalmers University of Technology.

Mark van der Laan

University of California, Berkeley

Statistical Inference with Gene Expression Data

We will provide a formal statistical framework for the analysis of gene expression. In particular, we consider estimation of clustering and subsetting parameters (genes and binding sites), estimation of the number of clusters, estimation of the reproducibility and reliability of subsets and clusters. We also consider the problem of prediction of an outcome such as survival based on gene expression data.

Ernst Wit

Department of Statistics, University of Glasgow

Hidden Markov Modelling of Genomic Expression Interactions

Introduction

Microarray technology has made the simultaneous measurement of gene transcription a routine activity. Whereas gene transcription is only one stage in the complex genomic process of living organisms, it gives a fascinating insight in one aspect of this activity across the whole genome. It is our aim in this paper to model local interaction behaviour in transcription for a tuberculosis bacterium in a stressed growth stage.

Gene regulation is a complex biological process which involves gene-gene and gene-protein interactions. The gene sequence is preceded by an operator region to which the enzyme polymerase is to bind to start transcription of the gene. However, many genes are normally blocked by the action of a repressor protein. This prevents the RNA polymerase enzyme from binding to the gene and transcribing the structural gene. Such genes are induced by the arrival of an inducer molecule which binds to the repressor protein and rendering it inactive. This allows transcription from the structural gene and the production of a protein.

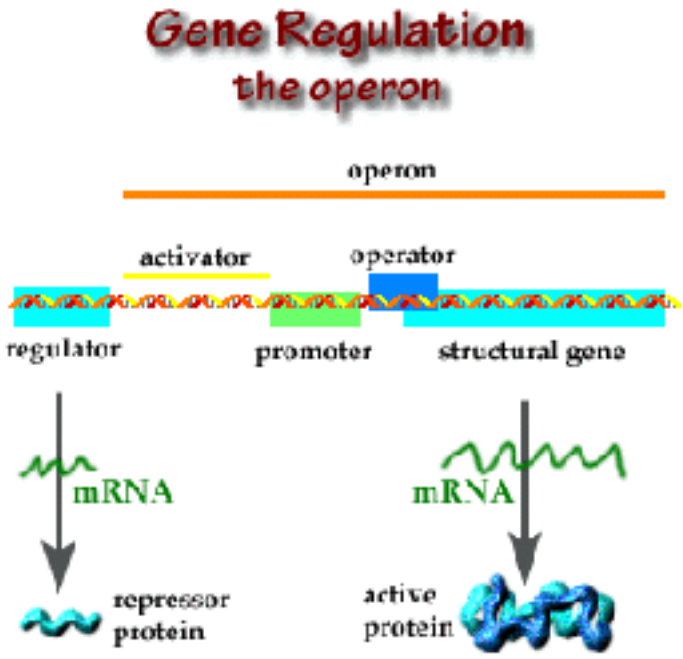

Figure 10: Transcription measured by microarrays is the end result of a complex series of biological regulatory operations.

Other genes are normally active and able to be constantly transcribed, because the repressor protein is produced in an inactive form. On the arrival and binding of the co-repressor molecule the complex can act as a functional repressor and block the structural gene by binding at the operator site.

In simple organisms genes are packed very close together. In fact, they frequently share a control region. Moreover, products from the transcription, i.e. proteins, can block the action of neighbouring genes. For this reason, it might not be unreasonable to test the hypothesis if there are some genome wide patterns of local spatial gene interactions.

Design of Experiment

In this section we describe an experiment conducted by Prof Phil Butcher and his Bacterial Microarray Group (Bμgs) at St. George's Hospital in London. The Bμgs group was interested in studying the effects of stressed growth on the expression levels of all 3924 genes in *Mycobacterium tuberculosis*.

From an initial time zero point 5 cultures of M. tuberculosis were grown. The cultures in the first two flasks were grown until day 6 and then harvested, whereas the others were grown and harvested at day 14, 20 and 30 respectively. For each time points four batches of RNA were extracted and three of them were labelled with Cy3-dCTP and one with Cy5-dCTP. 16 batches of genomic DNA were prepared and 12 of them were labelled with Cy5-dCTP and 4 with Cy3-dCTP. 16 mixtures of genomic DNA together with time course RNA were created and hybridized to the microarrays.

At the initial time point the bacterium has a lot of oxygen and is in a friendly growth environment. As the bacterium starts to multiply, it start to use up its resources. A lack of oxygen results and a stress reaction is induced in the bacterium. Effectively, growth slows down and eventually comes to a halt. Although the bacterium is grown in flasks, an analogous situation occurs in the human body, where the bacterium first multiplies quite quickly but then slows down. In the body, the bacterium is then eaten by microphages. In this state, they remain dormant in the body. Eventually they break out of the microphages for a final assault to the body, which is fatal.

The microarrays were spotted on whole genome arrays, produced by the Bμgs group. They were constructed by robotic spotting onto poly-lysine coated glass slides of PCR amplicons that ranged in length from 60 to 999 basepairs. The amplicons were derived from portions of each of the 3924 predicted open reading frames (ORF) of the sequenced strain of M. tuberculosis H37RV. Primer pairs were selected so to minimize cross-homology with all other ORF's.

Bayesian inference with hierarchical models

The time-series data from the tuberculosis growth experiment provide the opportunity to study the dynamic aspects of the gene expression patterns and possible relationships between genes. We focus on the question whether the dynamic structure of a gene's

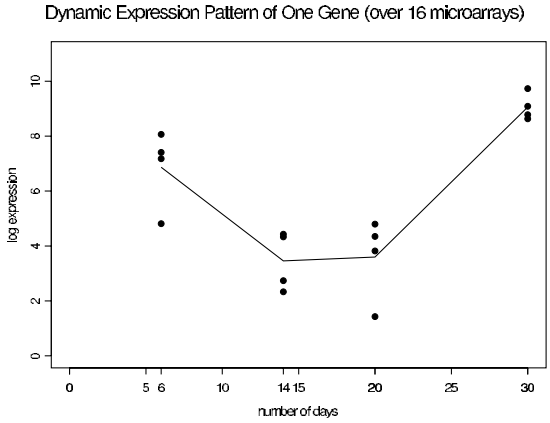

Figure 11: Down — Same — Up pattern for one gene.

expression pattern is related to the pattern of nearby genes on the genome. We model this by using a Hidden Markov Model interpreted as a Bayesian hierarchical model.

Current statistical work in microarray inference has rarely taken into account other aspects of the rich structure of genomic data. The sheer complexity of a microarray seems to have stunned the statistician into a pious acknowledgement of the exploratory nature of his or her contribution. This is a shame. It seems that the complete nature of the microarray datasets combined with knowledge of the structure of the genome allows for careful testing of interesting hypotheses. The main focus of this section is to test the spatial component of the M. tuberculosis gene expression measured during their growth stage.

Look at the data

A cursory look at the data shows something quite intriguing. The quantity,

$$\bar{x}_{it.} = \text{an average expression level for gene } i \text{ at time } t \tag{2.1}$$

is defined and calculated. From the four values $\{\bar{x}_{i1.}, \bar{x}_{i2.}, \bar{x}_{i3.}, \bar{x}_{i4.}\}$ an expression pattern is formed for gene i consisting of three signs $(+\backslash-)$ describing the increasing and/or decreasing pattern over time. When comparing the expression pattern of one gene with the next gene on the genome (where "next" is arbitrarily interpreted in one particular direction), it seems that a particular expression pattern is more likely to be followed by the inverse pattern.

From \ To	− − −	+ − −	− + −	+ + −	− − +	+ − +	− + +	+ + +
− − −	4	23	30	85	19	56	60	91
+ − −	8	41	60	58	63	74	152	86
− + −	27	64	50	90	59	97	84	65
+ + −	64	48	79	52	113	77	62	24
− − +	28	70	60	110	53	81	57	68
+ − +	56	97	103	60	65	40	64	27
− + +	87	145	95	51	66	55	35	9
+ + +	94	54	59	13	89	32	29	7

For instance, a strictly decreasing pattern is 91 times (25%) followed by a strictly increasing pattern, whereas a down-up-up pattern is followed 145 times (27%) by an up-down-down pattern, etc. This is quite a deviation from the expected 12.5% if there had been no spatial interaction. Nevertheless, this table is only a rough approximation of the true situation. The main problem with this table is that we expect most or at least a part of the genes to be unaffected by the growth phase. Any up-and-down pattern that is observed in such gene expressions is merely noise. Any serious modelling exercise should therefore allow for a "no-change" (0) as well as an "up" (+1) and "down" (−1) state.

Bayesian Hierarchical Model

The way we choose to model this problem is by a hierarchical model. Of interests are two things, i.e., the unobserved dynamic expression patterns made up out of "up", "down" and "no-change" and neighbour interactions patterns between these unobserved states.

We define the hidden states $s = (s_1, \ldots, s_{3924})$ on the genes in the tuberculosis genome. Each hidden state variable is a collection of 4 subvariables, $s_1 = (s_{i1}, s_{i2}, s_{i3}, R_i)$, where $s_{ij} \in \{-1, 0, +1\}$ expresses whether gene i from time instance j to $j+1$, went down, stayed the same or went up. The parameter R_i expresses to which interaction regime gene i belongs. For the course of this paper we assume that the interaction regime is homogeneous over the whole genome. It is known that the M. tuberculosis genome has a circular structure, as shown in Figure 12. The 3924 known genes in this genome have the peculiar one-dimensional structure, whereby gene nr. 3924 is located right next to gene nr. 1.

To simplify the modelling, we shall concentrate only on the observed average differences with respect to the previous time point,

$$d_{i1} = \bar{x}_{i2.} - \bar{x}_{i1.}, \quad d_{i2} = \bar{x}_{i3.} - \bar{x}_{i2.}, \quad d_{i3} = \bar{x}_{i4.} - \bar{x}_{i3.},$$

Given the unobserved states, we shall model the differences with distributions depending on a parameter θ_o.

The structure of the hierarchical model reflects the circularity of the genome. In Figure 13 the observed differences $d_i = (d_{i1}, d_{i2}, d_{i3})$ for $i = 1, 2, \ldots 3924$ are shown in their dependence to the observation parameters θ_o as well as on the circular structure of the genome. The hidden state parameters s depend on some parameters θ_m.

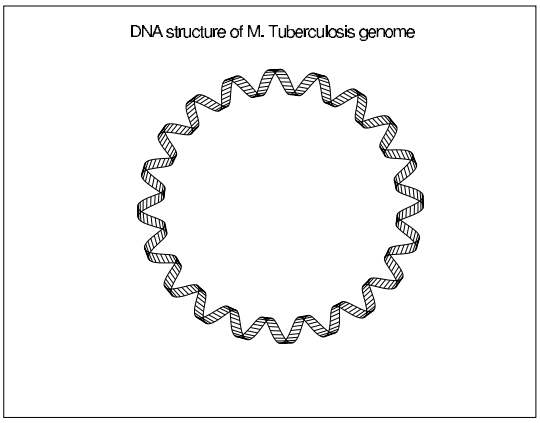

Figure 12: The 3924 genes on the M. Tuberculosis genome are arranged in a circular fashion.

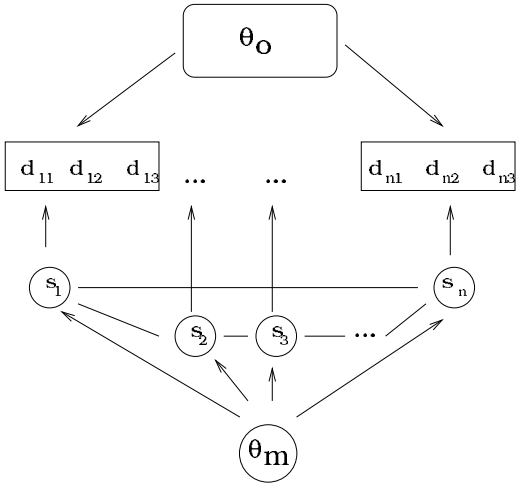

Figure 13: A graphical representation of the tuberculosis spatial interaction model. Notice that this is not a DAG. The cyclic structure of the genome introduces a cyclic structure among the variables.

Hidden Potts Model

One of the questions of interest is to compare the independence model where the gene expression pattern of gene i, s_i, does not depend on that of neighbouring genes s_{i-1} and s_{i+1} with more complicated spatial models. Implementing the hierarchical Bayesian model as a hidden Potts model gives us precisely this possibility.

In general, if one interprets the circular genome as a conditional independence graph, the Hammersley-Clifford theorem (Besag 1974) suggests that the full prior distribution on the genome $s = (s_1, s_2, \ldots, s_{3924})$ is given as

$$P(s) \propto \exp\left(\sum_i \psi_i(s_i, s_{i+1})\right).$$

This first order model only takes neighbouring effects into consideration. We choose a particular interpretation of this model, by means of a Potts model.

The Potts model defines the joint distribution through the conditional distributions. It is a generalization of an Ising model, where each of the positions can take on more than two levels, in this case three, i.e., $\{-1, 0, +1\}$. The Markov Random Field in our problem is two-dimensional, where the first dimension comes from the spatial structure of the genome and the other dimension from time.

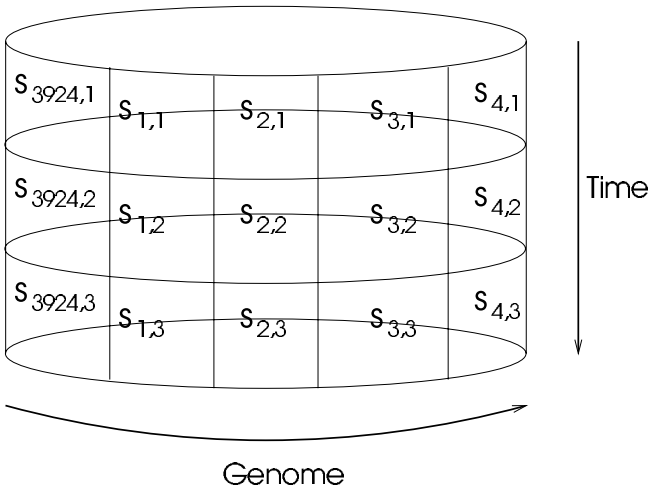

Figure 14: The hidden parameter s defines a Markov Random Field on a lattice that wraps around. The hidden Markov Random Field is modelled through a Potts Model.

The Potts model is defined through the conditional probabilities, $p(s_{ij}|s_{-ij}\theta_m)$. Using the Markov structure by which these conditional probabilities only depend on the states within the neighbourhood structure, a model can be defined with a richness of one's own liking. We define the following Potts model with six parameters $\theta_m = (\theta_{000}, \theta_{001}, \theta_{010}, \theta_{011}, \theta_{100}, \theta_{101})$,

$$
\begin{aligned}
&p(s_{i_0j_0}|s_{-i_0j_0}\theta_m) \propto \qquad (2.2)\\
&\quad \exp\Big(\theta_{000}1_{\{s_{i_0j_0}=0\}} + \theta_{001}1_{\{s_{i_0j_0}=1\}}\\
&\quad +\theta_{100}\sum_{i\in\{i_0-1,i_0+1\}} 1_{\{s_{i_0j_0}=s_{ij_0}=0\}} + \theta_{101}\sum_{i\in\{i_0-1,i_0+1\}} 1_{\{s_{i_0j_0}=s_{ij_0}\}}\\
&\quad + \theta_{010}\sum_{j\in\{j_0-1,j_0+1\}} 1_{\{s_{i_0j_0}=s_{i_0j}=0\}} + \theta_{011}\sum_{j\in\{j_0-1,j_0+1\}} 1_{\{s_{i_0j_0}=s_{i_0j}\}}\Big)
\end{aligned}
$$

The first two parameters guarantee that if $\theta_{010} = \theta_{011} = \theta_{100} = \theta_{101} = 0$, the model on the hidden states reduces to full independence in both the time and spatial dimensions. If the parameter $\theta_{101} > 0$, then then genes tend to keep their state from one time to the next. If the parameter $\theta_{101} < 0$ then genes tend to flip to another state from one time to the next. The parameter θ_{100} focusses particularly on the behaviour of non-expressed genes over time. If the parameter $\theta_{011} > 0$, then then neighbouring genes tend to have the same state. If the parameter $\theta_{011} < 0$ then neighboring genes tend to repulse each other. The

parameter θ_{010} focusses particularly how the non-active behavious of one gene propogates to genes neighbouring it.

The full conditionals in (2.2) determines the joint distribution of the hidden states $p(s|\theta_m)$. It can be shown with the use of the factorization theorem in probabilistic graphical models that the joint distribution of the hidden states given the model parameters θ_m is proportional to the following expression,

$$\begin{aligned} p(s|\theta_m) \propto & \\ & \exp\left(\theta_{000}\sum_{i_0j_0} 1_{\{s_{i_0j_0}=0\}} + \theta_{001}\sum_{i_0j_0} 1_{\{s_{ij}=1\}}\right. \\ & +\theta_{100}\sum_{i_0j_0} 1_{\{s_{i_0j_0}=s_{i_0+1,j_0}=0\}} + \theta_{101}\sum_{i_0j_0} 1_{\{s_{i_0j_0}=s_{i_0+1,j_0}\}} \\ & \left. + \theta_{010}\sum_{i_0j_0} 1_{\{s_{i_0j_0}=s_{i_0,j_0+1}=0\}} + \theta_{011}\sum_{i_0j_0} 1_{\{s_{i_0j_0}=s_{i_0j_0+1}\}}\right), \end{aligned} \tag{2.3}$$

where the sums are over the whole lattice and where for all i the quantity s_{i4} is simply defined as missing.

We assume that the unobserved states s determine the distribution of the average difference between the time points. In particular, we assume that given a particular gene i in state $s_i = (s_{i1}, s_{i2}, s_{i3})$ the observed differences are distributed as correlated normals,

$$d_i|s_i \sim N\left(\begin{pmatrix} \mu_{s_{i1}} \\ \mu_{s_{i2}} \\ \mu_{s_{i3}} \end{pmatrix}, \begin{pmatrix} \sigma^2_{s_{i1}} & -\sigma_{s_{i1}}\sigma_{s_{i2}}/2 & 0 \\ -\sigma_{s_{i1}}\sigma_{s_{i2}}/2 & \sigma^2_{s_{i2}} & -\sigma_{s_{i3}}\sigma_{s_{i2}}/2 \\ 0 & -\sigma_{s_{i3}}\sigma_{s_{i2}}/2 & \sigma^2_{s_{i3}} \end{pmatrix}\right)$$

under the restriction that

$$\mu_{-1} < 0, \quad \mu_0 \equiv 0 \quad \text{and} \quad \mu_1 > 0.$$

Given the states s the observed differences are assumed to be independent.

Computational Details of Implementation

The set-up of the model falls completely within the hidden Markov description by [**?**] with the additional computational advantages of [**?**]. The purpose of inference is to get an estimate of the posterior $p(s, \theta_o, \theta_m \mid d)$. This estimation of the parameters is done via a hybrid sampler.

We let $(d, s) = \{(d_i, s_i), i = 1, \ldots, n\}$, where $n = 3924$ and for each gene i is the set of observed differences corresponding to an unobserved missing dynamic state s_i. Conditional on the s_i's, the d_i's are independent,

$$p(d|s, \theta_0, \theta_m) = \prod_{i=1}^{n} p(d_i|s_i, \theta_o, \theta_m).$$

The conditional independence graph in Figure 13, in particular the independence of θ_o and θ_m, allows the following factorization

$$p(d, s, \theta_o, \theta_m) = \left\{\prod_{i=1}^{n} p(d_i|s_i, \theta_o)\right\} p(s|\theta_m)\pi_1(\theta_m)\pi_2(\theta_o), \tag{2.4}$$

where π_1 and π_2 are priors for θ_m and θ_o, respectively.

First, let s_{-i} denote $\{s_j, j \neq i\}$. Then we may write

$$p(s|\theta_m) = p(s_i|s_{-i}\theta_m)p(s_{-i}|\theta_m).$$

Furthermore, let $s_{\delta i}$ denote the subset of s_{-i} that contribute to $p(s_i|s_{-i}\theta_m)$. In other words, $s_{\delta i}$ contains the *neighbours* of s_i. In the case of our circular Markov random field, $s_{\delta i} = \{s_{i-1}, s_{i+1}\}$, where $s_0 \equiv s_{3924}$ and $s_{3925} \equiv s_1$.

Using this notation, a Gibbs sampler can be defined to generate draws from $p(s, \theta_o, \theta_m \mid d)$. After a suitable burn-in period, draws for s, θ_o and θ_m can be obtained iteratively by simulating according to the following cycle:

1. For each i, simulate s_i from

$$\begin{aligned} p(s_i|s_{-i}, \theta_o, \theta_m) &= p(s_i|s_{\delta i}, \theta_o, \theta_m) && (2.5)\\ &\propto p(d_i|s_i, \theta_o)p(s_i|s_{\delta i}, \theta_m) && (2.6) \end{aligned}$$

   Remember that s_i is a categorical variable, and therefore sampling from $p(s_i|s_{-i}, \theta_o, \theta_m)$ is effectively sampling from several multinomials over time.

2. Sample θ_o from

$$p(\theta_o|\theta_m, s, d) \propto \left\{\prod_{i=1}^{n} p(d_i|s_i, \theta_o)\right\} \pi_2(\theta_o) \tag{2.7}$$

   By selecting a conjugate prior, evaluating (2.7) is straightforward.

3. Draw θ_m from

$$\begin{aligned} p(\theta_m|\theta_o, s, d) &= p(\theta_m|s) && (2.8)\\ &\propto p(s|\theta_m)\pi_1(\theta_m) && (2.9) \end{aligned}$$

   The expression $p(s|\theta_m)$ is proportional to the expression in (2.3). It is known therefore up to a normalizing constant $Z(\theta_m)$ that depends on θ_m.

Whereas the parameters s and θ_o are easy to update via a Gibbs or other Hasting scheme, there is a problem when it comes to updating θ_m. The fact that the expression (2.9) depends on the normalizing constant of the joint density of the hidden states, has posed obstacles to sample θ_m effectively. [**?**] use the pseudolikelihood,

$$p_{PL}(s|\theta_m) = \prod_i p(s_i|s_{\delta i}\theta_m),$$

as an approximation of $p(s|\theta_m)$ and replace this in (2.9).

A recent work by [**?**] found an important result for efficient and precise calculation of the normalizing constant $z(\theta_m)$. Let $A = \{a_1, \ldots, a_N\}$ be the set of values s_i can take with cardinality $|S| = N$. It states that on a cylindrical lattice s the normalizing constant can be found for a non-normalized joint density q that can be factorized as

$$q(s|\theta) = \prod_{j=1}^{n} h(s_j, s_{j-1}),$$

for a given positive real function $h : A \times A \to R$. The normalizing constant is given by $\mathrm{tr}(Q^n)$, where Q is a $N \times N$ matrix, defined by

$$Q_{kl} = h(x_j = a_l, x_{j-1} = a_k).$$

In our case, the set A of values that s_i can take consists of $3^3 = 27$ elements,

$$A \;=\; ((000), (00+), (00-), \ldots, (+++)) \,. \tag{2.10}$$

From the joint distribution in (2.3) it is easy to see that the function h can be written as

$$\begin{aligned} h(s_i, s_{i+1}) \;=\; \exp\Bigg(\sum_{j=1}^{3} \Big(&\theta_{000} 1_{\{s_{ij}=0\}} + \theta_{001} 1_{\{s_{ij}=1\}} + \theta_{100} 1_{\{s_{ij}=s_{i+1,j}=0\}} \\ &+\theta_{101} 1_{\{s_{ij}=s_{i+1,j}\}} + \theta_{010} 1_{\{s_{ij}=s_{i,j+1}=0\}} + \theta_{011} 1_{\{s_{ij}=s_{i,j+1}\}}\Big)\Bigg) \,, \end{aligned}$$

where for each i the value s_{i4} in this expression is assumed to be missing. From this we can derive the 27×27 matrix Q. Using the order as partially defined by (2.10) we can write for Q_L, the elementwise log-transformation of Q:

$$Q_L \;=\; \begin{pmatrix} 3\theta_{000} + 3\theta_{100} + 3\theta_{101} + 2\theta_{010} + 2\theta_{011} & \ldots & 3\theta_{000} + 2\theta_{010} + 2\theta_{011} \\ 2\theta_{000} + \theta_{001} + 2\theta_{100} + 2\theta_{101} + \theta_{010} + \theta_{011} & \ldots & 2\theta_{000} + \theta_{001} + \theta_{101} + \theta_{010} + \theta_{011} \\ 2\theta_{000} + 2\theta_{100} + 2\theta_{101} + \theta_{010} + \theta_{011} & \ldots & 2\theta_{000} + \theta_{010} + \theta_{011} \\ \vdots & \ddots & \vdots \\ 3\theta_{001} + 2\theta_{011} & \ldots & 3\theta_{001} + 3\theta_{101} + 2\theta_{011} \end{pmatrix}$$

The matrix Q consists of strictly positive entries and is therefore irreducible. The Perron-Frobenius matrix theorem applies so that Q can be diagonalised, $Q = H^{-1}DH$. Then $\mathrm{tr}(Q^n) = \mathrm{tr}(D^n)$, which gives substantial savings in terms of calculation.

Discussion and Results

Currently the a Hybrid Gibbs and Metropolis-Hasting sampler, outlined in Section 3, is implemented. Initial results suggest:

1. Using the exact computation method described in Section 3.4. is computationally as expensive as the traditional pseudo-likelihood approach, whereas the efficiency of the results is much greater.

2. The peculiar patterns observed in the up-down patterns in the raw data are confirmed with our formal model.

Although our model is very naive and should be supplemented by more detailed modelling, it is one of the first attempts to look at expression data from a modelling point of view. Moreover, although the biological story is no doubt much more complicated, the current results suggest that some form of localized negative feedback might be part of the regulatory activity of the M. tuberculosis bacterium's genome.

Torben F. Ørntoft

Dept. Clinical Biochemistry, Aarhus University Hospital, Skejby, DK-8200 Aarhus N, Denmark. Ornttoft@kba.sks.au.dk

Microarrays, basic principle and use in medical research

Basic principle

The immobilization of probes for nucleic acid hybridization on a solid support has been known for many years, but the technique was taken to a new level when thousands of molecules were placed next to each other on a glass surface in a small microscopic area with less than 100 micrometers between the probes. This enabled the use of small volumes of samples to be hybridized as well as parallel analysis of thousands of genes - today reaching the level of examining the whole transcribed genome in one step on approximately 1 square centimeter of glass.

The perspective for this approach was and still is staggering. Placing 400.000 probes on a small glass surface makes it possible to design probes for sequencing, for polymorphism detection, for expression, splice variant examination etc.

The most widely used commercial array platform is from Affymetrix Inc. and involves 25-mer oligonucleotides synthesized in-situ on a solid glass wafer surface by a photolithographic process. Approximately 20 probes are selected for each gene placed towards the 3' end of the ORF and with matching control probes for control of unspecific cross hybridization. These arrays are for single sample measurements that are normalized and scaled for comparison to other samples. Non-commercial platforms usually utilize robotic gridders or inkjet printing technology that deposit the DNA probes on the arrays surface. Reading is made in a laser scanning confocal microscope with a resolution around 2-5 microns or with CCD camera technology.

Application in Gene expression monitoring

The most widespread application of arrays has, by far, been the quantification of mRNA molecules, so-called gene expression monitoring or transcriptome analysis. In this application information about which genes are expressed and at what level is provided. The performance is linear and reproducible and an antibody sandwich similar to an ELISA is used to amplify the signal from weakly expressed sequences Another approach has been sequencing in which probes that ideally cover all 4 possible nucleotide positions along the whole open reading frame are used to predict the presence or absence of mutations (Wikman et al., 2000). A fast-growing application is the SNP array-technology with the purpose of detecting variation in the genome to find polymorphisms associated to increased disease susceptibility or to e.g. reduced or increased metabolism of drugs (Halushka et al., 1999, Sekine et al, 2001). About 4-6 millions of polymorphisms exist in the genome, the vast amount outside of open reading frames.

The sample that is loaded on the expression arrays is labeled following one of several protocols. Nano- to micrograms of total RNA is needed and labeled by an enzymatic

reaction or using fluorescent labeled or modified nucleic acids. In the two most used protocols 5-20 μg total RNA is used to synthesize cDNA by reverse transcription or cRNA by in vitro transcription from a double-stranded DNA modified with a 3'- T7 RNA polymerase recognition site. Non-commercial systems normally incorporate fluorescent Cy5- or Cy3-conjugated nucleotides in test- and reference samples, respectively, or amino-allyl nucleotides that are subsequently conjugated with cy-dyes. Following the Affymetrix protocol, biotin-modified ribonucleotides are incorporated during in vitro transcription and reacted with a streptavidin-phycoerythrin conjugate after hybridization. Subsequently, an antibody sandwich similar to an ELISA is used to amplify the signal in order to detect weakly expressed sequences.

Modified protocols exist in which several rounds of in vitro transcription are leading to detectable amounts of labeled cRNA from a few nanograms of RNA. One approach is not to use PCR but only approximately linear enzymatic amplification, whereas others with success have used a PCR step (Theilgaard-Monch et al., 2001).

Handling of data

Following an array experiment on the Affymetrix system the initial scan has to be scaled to a predefined standard level. This is necessary to compensate for e.g. varying efficiency during labelling of the samples. For this reason it is always wise to label samples for comparison in parallel on the same day. The standardization can be based on a global scaling in which the intensity from all probes on the array is scaled to a standard level or can be based on the level of selected house-keeping genes that are supposedly at the same level in different samples under different conditions. A list of such genes is available at (Warrington et al., 2000). If the scaling factors between the samples to be compared differs more than approximately three fold, the validity of the comparison becomes uncertain.

The next step is bioinformatic datamining with the purpose of characterizing the samples or identifying genes whose expression is following a certain sample phenotype. There are various approaches to this.

First step is to eliminate genes whose expression is generally absent or noise-filled. Methods for this depend on the array system used. The second step is to identify those genes whose expression is informative in terms of following a certain sample phenotype or in terms of having a differential behavior across the samples. This step will eliminate a large number of genes that do not vary much as e.g. housekeeping genes. Different methods exist for this purpose, some of which are,

1. Statistical evaluation based on parametric statistics like t-test. In this case the expression of a gene across many samples is normalized and statistically significant variation is registered.

2. Using a weighting scheme in which genes whose expression covary with a certain sample phenotype are sorted for further study.

3. Significance analysis of microarrays. This method assigns a score to each gene based on changes in gene expression relative to the standard deviation of repeated measurements (Tusher et al., 2001).

When the differentially expressed genes have been identified, the third step is to extract

as much information from the informative pool of genes as possible. A commonly used method for this purpose is cluster analysis based on hierarchical agglomerative clustering in which similar gene expression levels across samples lead to a tight relation between genes and a dissimilar behavior leads to clustering of such genes far from each other. The same method can be applied to samples (two-way clustering) such that samples with similar expression of a number of genes are clustered together and those that have different levels of certain groups of gene are clustered far from each other. In this way a new gene expression based classification of samples occur, unbiased from other assumptions of the samples. This has been very promising in identifying new subgroups of diseases based on gene expression. Such subgroups do in cases correspond to patients with good or poor survival (Alizadeh et al., 2000, van 't Veer et al., 2002), patients who will benefit from certain treatments etc. As many of these projects start out with thousands of genes it is perhaps not surprising that interesting groups of for example 70 genes that work as classifiers will be found. The crucial point is whether such classifying genes will reproducibly identify the same class of patients or samples in a new independent experiment. Mathematical approaches as cross-validation "within" the experiment are made to support the robustness of the classifier, but cannot substitute prospective clinical studies that are quite time consuming.

A promising application of expression arrays has been the identification and dissection of regulatory pathways. If a certain receptor is triggered by a ligand or if an inducible signal transduction molecule is activated following transfection (Pedersen et al., 2001) or physical stimulation (Zhao et al., 2000), a very clear impression of the effect on gene transcription can be monitored. New software allows the direct linkage of expression data to pathways for easy visualization. This fast and comprehensive overview of large datasets of gene expression profiles may lead to a faster discovery of new pathways and interaction between pathways. In a similar manner knockout or knock-in mice can be used to study the biological role of specific genes.

Application in SNP monitoring

SNP arrays are mainly used for two purposes, I) linkage analysis of certain chromosomal loci identified by SNPs to diseases for positional cloning or to find the genetic background of patients' response to drugs (Sekine et al., 2001; Halushka et al., 1999); II) detection of allelic imbalance in tumor tissue (Primdahl et al., 2002). In the first case large clinically well-defined family cohorts are needed, and regions segregating with disease are subjected to fine mapping with microsatellites, database search for gene identification and sequencing of candidate genes. In the latter case DNA purified from micro-dissected tumor tissue is compared to the normal leukocyte DNA, and heterogeneous SNP's will convert to homozygous if the allelic balance is changing. The method can at least in theory detect both losses and gains of alleles, but mostly lost alleles are picked up, probably due to saturation problems in the PCR reaction that has to be carried out as heavy multiplexing (more than 50 primer pairs per tube) to reduce the workload. The PCR step in which a small oligonucleotide holding the SNP's has to be amplified seems for the time being to be the rate limiting step in the process towards SNP arrays with more than 10.000 SNPs.

The use of arrays for sequencing has been hampered by the fact that the sensitivity is

not as high as with conventional acrylic polymer based sequencers and identification of deletions and insertions is at present very difficult. Data indicate that each probe has to be characterized with respect to signal to noise ratio etc. This makes it necessary to run a very large amount of controls before utilizing these arrays (Wikman et al 2000). The many databases created worldwide using array systems, are potentially rich sources of information. With regard to the Affymetrix system, that should be straight forward, however, other arrays systems are very difficult to compare as the linearity and dynamics may vary considerably not to mention the references (pools of RNA) that are used for comparisons (Brazma et al., 2001). We will probably need standardized commercial or publicly distributed references to utilize the potential of the many data worldwide that describe different species, tissues and biological conditions.

References

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503-11

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29:365-71

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22:239-47

Hitoshi Zembutsu, Yasuyuki Ohnishi, Tatsuhiko Tsunoda, Yoichi Furukawa, Toyomasa Katagiri, Yoshito Ueyama, Norikazu Tamaoki, Tatsuji Nomura, Osamu Kitahara, Rempei Yanagawa, Koichi Hirata and Yusuke Nakamura (2002). Genome-wide cDNA Microarray Screening to Correlate Gene Expression Profiles with Sensitivity of 85 Human Cancer Xenografts to Anticancer Drugs. Cancer Research 62, 518-527

Pedersen MW, Thykjaer T, Orntoft TF, Damstrup L, Poulsen HS (2001). Profile of differentially expressed genes mediated by the type III epidermal growth factor receptor mutation expressed in a small-cell lung cancer cell line. Br J Cancer 85:1211-8

Primdahl H, Wikman FP, von Der Maase H, Zhou Xg X, Wolf H, Orntoft TF. (2002). Allelic Imbalances in Human Bladder Cancer: Genome-Wide Detection With High-Density Single-Nucleotide Polymorphism Arrays. J Natl Cancer Inst 94 : 216-223

Thykjaer T, Workman C, Kruhoffer M, Demtroder K, Wolf H, Andersen LD, Frederiksen CM, Knudsen S, Orntoft TF (2001). Identification of gene expression patterns in superficial and invasive human bladder cancer. Cancer Res 61:2492-9

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature 415 :530-6

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES, et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077-82

Wikman FP, Lu ML, Thykjaer T, Olesen SH, Andersen LD, Cordon-Cardo C, Orntoft TF (2000). Evaluation of the performance of a p53 sequencing microarray chip using 140 previously sequenced bladder tumor. Clin Chem 46 :1555-61

Zhao R, Gish K, Murphy M, Yin Y, Notterman D, Hoffman WH, Tom E, Mack DH, Levine AJ (2000). Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. Genes Dev 4:981-93

Sekine A, Saito S, Iida A, Mitsunobu Y, Higuchi S, Harigae S, Nakamura Y (2001). Identification of single-nucleotide polymorphisms (SNPs) of human N-acetyltransferase genes NAT1, NAT2, AANAT, ARD1 and L1CAM in the Japanese population. J Hum Genet 46:314-9.

Theilgaard-Monch K, Cowland J, Borregaard N (2001). Profiling of gene expression in individual hematopoietic cells by global mRNA amplification and slot blot analysis. J Immunol Methods 252:175-89

Warrington, JA., Nair, A., Mahadevappa, M., Tsyganskaya, M (2000). Comparison of human adult and fetal expression and identification of 535 housekeping / maintenance genes. Physiological Genomics 2:143-147.

Tusher VG, Tibshirani R, Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98:5116-21.

3 List of participants

Trygve Almøy
Department of Mathematical Sciences
P.O. Box 5035
N-1432 Aas
Norway
Email: trygve.almoy@imf.nlh.no

Alessandro Ambrosi
Department of Oncological Sciences
Cl.CH.II
Via Giustiniani, 2
35128 Padua, Italy
Email: ale.ambrosi@unipd.it

Elisabeth W. Andersen
Retsgenetisk Afdeling
Københavns Universitet
Frederik Vs vej 11
2100 Copenhagen Ø, Denmark
Email: Elisabeth.Andersen@forensic.ku.dk

Konstantin Arbeev
Max Planck Institute for Demographic Research
Laboratory of Advanced Statistical Methods
Konrad-Zuse-Str. 1
D-18057 Rostock, Germany
Email: arbeev@demogr.mpg.de

Nicola Armstrong
EURANDOM
PO Box 513
5600 MB Eindhoven
The Netherlands
Email: armstromg@eurandom.tue.nl

Jörg Assmus
Matematisk Institut
Johannes Bruns Gate 12
N-5008 Bergen
Norway
Email: assmus@mi.uib.no

Jens Henrik Badsberg
Danmarks Jordbrugsforskning
Forskningscenter Foulum
8830 Tjele
Denmark
Email: JensHenrik.Badsberg@agrsci.dk

Søren Bak
The Royal Veterinary and Agricultural University
Department of Plant Biology
Thorvaldsensvej 40, opgang 6, sal 2
1871 Frederiksberg C, Denmark
Email: bak@kvl.dk

Bojan Basrak
EURANDOM
PO Box 513
5600 MB Eindhoven
The Netherlands
Email: basrak@eurandom.tue.nl

Viktor Benes
Department of Probability and Statistics
Charles University, Faculty of Mathematics and Physics
Sokolovska 83
18675 Praha 8, Czech Republic
Email: benesv@karlin.mff.cuni.cz

Alessandra Rosalba Brazzale
Institute for Biomedical Engineering
Italian National Research Council
Corso Stati Uniti 4
35127 Padova, Italy
Email: alessandra.brazzale@isib.cnr.it

Lisbeth Carstensen
Inst. for Folkesundhedsvidenskab
Panum Instituttet
Blegdamsvej 3
2200 Copenhagen N, Denmark
Email: lisbeth@rhk.dk

Giovanna Chiorino
Fondo Edo Tempia
via Malta 3
13900 Biella
Italy
Email: giovanna.chiorino@fondoedotempia.it

Ole Christensen
BiRC - Bioinformatics Research Center, Dept. of Computer Science
University of Aarhus
Ny Munkegade
DK-8000 Aarhus C, Denmark
Email: o.christensen@lancaster.ac.uk

Peter Dalgaard
Department of Biostatistics
University of Copenhagen
Blegdamsvej 3
2200 Copenhagen N, Denmark
Email: p.dalgaard@biostat.ku.dk

Mathisca de Gunst
Department of Mathematics
Vrije Universiteit
De Boelelaan 1081a
1081 HV Amsterdam, The Netherlands
Email: degunst@cs.vu.nl

Joan del Castillo
Department of Mathematics
Campus de la UAB, Facultat de Ciencies
08193 Cerdanyola del Valles
Spain
Email: castillo@mat.uab.es

Susanne Ditlevsen
Department Biostatistics
University of Copenhagen
Blegdamsvej 3
2200 Copenhagen N, Denmark
Email: sudi@pubhealth.ku.dk

Lars Dyrskjøt
Molecular Diagnostic Laboratory
Aarhus University Hospital
Skejby
8200 Aarhus N, Denmark
Email: lars@kba.sks.au.dk

David Edwards
Department of Biostatistics
Novo Nordisk
DK-2880 Bagsværd
Denmark
Email: ded@novonordisk.com

Magnus Ekdahl
Linköpings Universitet
Fanjunkaregatan 45
53228 Linköping
Sweden
Email: maekd@mai.liu.se

Jacob Engelbrecht
Microbial Cell Technology
Novo Nordisk A/S
6A1.038 Novo Alle
DK-2880 Bagsværd, Denmark
Email: jaen@novonordisk.com

Guri Feten
Department of Mathematical Sciences
P.O. Box 5035
N-1432 Aas
Norway
Email: guri.feten@imf.nlh.no

Barbel Finkenstadt
Department of Statistics
University of Warwick
CV4 7AL Coventry
United Kingdom
Email: stsbe@csv.warwick.ac.uk

Arnoldo Frigessi
Norwegian Computing Center
P.O.Box 114 Blindern
N-0314 Oslo
Norway
Email: frigessi@nr.no

Jelle Goeman
Medical Statistics
Leiden University Medical Center
2300 RC Leiden
The Netherlands
Email: j.j.goeman@lumc.nl

Jørgen Granfeldt
Department of Mathematical Sciences
University of Aaarhus
Ny Munkegade
DK-8000 Aarhus C, Denmark
Email: jqrgen@imf.au.dk

Per Gregersen
Danish Institute of Agricultural Sciences
Department of Plant Biology
Forskningscenter Flakkebjerg
DK-4200 Slagelse, Denmark
Email: per.gregersen@agrsci.dk

Peter Hagedorn
Risø National Laboratory
Plant Research Department, Building 776
Frederiksborgvej 399
4000 Roskilde, Denmark
Email: hagedorn@nbi.dk

Ernst Hansen
Afdeling for Anvendt Matematik og Statistik
Københavns Universitet
Universitetsparken 5
2100 Copenhagen Ø, Denmark
Email: erhansen@math.ku.dk

Kasper Daniel Hansen
Department of Biostatistics
University of Copenhagen
Blegdamsvej 3
2200 Copenhagen N, Denmark
Email: K.Hansen@biostat.ku.dk

Svend Erik Westh Hansen
DakoCytomation Denmark A/S
Produktionsvej 42
DK-2600 Glostup
Denmark
Email: sebra@post6.tele.dk

Esben Hedegaard
The Statistics Division
University of Copenhagen
2100 Copenhagen
Denmark
Email: esben@hedegaard.name

Anne-Mette Hein
Department of Epidemiology and Public Health
St. Mary's Medical School
Imperial College, Norfolk Pl.
W2 1PG London, United Kingdom
Email: a.hein@ic.ac.uk

Sonia Hernandez
EURANDOM
P.O. Box 513
5600 MB Eindhoven
The Netherlands
Email: hernandez@eurandom.tue.nl

Charlotte Hindsberger
Department of Biostatistics
Blegdamsvej 3
2200 Copenhagen N
Denmark
Email: chh@biostat.ku.dk

Michael Höhle
Biometric Research Unit
Danish Institute of Agricultural Sciences
Research Centre Foulum
DK-8830 Tjele, Denmark
Email: hoehle@dina.kvl.dk

Fabian Hoti
Rolf Nevanlinna Institute
University of Helsinki
P.O.Box 4
FIN-00014 University of Helsinki, Finland
Email: fjh@rni.helsinki.fi

Ivan Iachine
Department of Statistics
University of Southern Denmark
Campusvej 55
5230 Odense M, Denmark
Email: iachine@statdem.sdu.dk

Stefano Iacus
Department of Economics
Via Conservatorio, 7
I-20122 Milan
Italy
Email: stefano.iacus@unimi.it

Ulf Indahl
Department of Mathematical Sciences
P.O. Box 5035
N-1432 Aas
Norway
Email: ulf.indahl@imf.nlh.no

Rafael Irizarry
Department of Biostatistics
Johans Hopkins University
615 N. Wolfe St. E3035
Baltimore, MD 21205, USA
Email: rafa@jhu.edu

Pretim Jain
Inst. of Genomics and Integrative Biology
132, c-7, sector 7
Rohini
110085 NewDelhi, India
Email: preti_jain1@yahoo.co.in

Daniel Jeffares
Department of Evolutionary Biology
University of Copenhagen
Universitetsparken 15
2100 Copenhagen Ø, Denmark
Email: dcjeffares@zi.ku.dk

Eva Vedel Jensen
Department of Mathematical Sciences
University of Aaarhus
Ny Munkegade
DK-8000 Aarhus C, Denmark
Email: eva@imf.au.dk

Jens Ledet Jensen
Department of Mathematical Sciences
University of Aaarhus
Ny Munkegade
DK-8000 Aarhus C, Denmark
Email: jlj@imf.au.dk

Mads Aaboe Jensen
Department of Clinical Biochemistry
Aarhus University Hospital
Skejby
DK-8200 Aarhus N, Denmark
Email: mads.aaboe@kba.sks.au.dk

Hans Arnfinn Karlsen
Department of Mathematics
University of Bergen
Johannes Bruns gt. 12
N-5008 Bergen, Norway
Email: karlsen@mi.uib.no

Niels Keiding
Department of Biostatistics
University of Copenhagen
Blegdamsvej 3
2200 Copenhagen N, Denmark
Email: N.Keiding@biostat.ku.dk

Steen Knudsen
Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark
Email: steen@cbs.dtu.dk

Timo Koski
Matemtiska Institutionen
Linköpings Universitet
SE-581 83 Linköping
Sweden
Email: tikos@mai.liu.se

Anders Krogh
Bioinformatik Centret
University of Copenhagen
Universitetsparken 15
2100 Copenhagen Ø, Denmark
Email: krogh@binf.ku.dk

Mogens Kruhøffer
Molecular Diagnostic Laboratory
Aarhus University Hospital¡BR¿Skejby
DK-8200 Aarhus N, Denmark
Email: mkr@iekf.au.dk

Corrado Lagazio
Department of Statistics
University of Udine
Via Treppo 18
33100 Udine, Italy
Email: lagazio@dss.uniud.it

Mette Langaas
Department of Mathematical Sciences
Norwegian University of Science and Technology
NO-7491 Tronheim
Norway
Email: mette.langaas@math.ntnu.no

Björn Larsson
Deparment of Mathmatics
Linköping University
S-581 83 Linköping
Sweden
Email: bjlar@mai.liu.se

Volkmar Liebscher
Institute of Biomathematics and Biometry
GSF-National Research Centre for Environment and Health
85764 Neuherberg / München
Germany
Email: liebscher@gsf.de

Patrick Lindsey
EURANDOM
PO Box 513
5600 MB Eindhoven
The Netherlands
Email: plindsey@euridice.tue.nl

Mogens Lund
Danish Institute of Agricultural Sciences
Inst. of Animal Breeding and Genetics
Vestergade 7
DK-8830 Tjele, Denmark
Email: Mogens.Lund@agrsci.dk

Tristan Mary-Huard
Institut National Agronomique Paris-Grignon
16 rue Claude Bernard
F-75231 Paris cedex 05
France
Email: maryhuar@inapg.fr

Claus-D. Mayer
Biomathematics & Statistics Scotland
Rowett Research Institute
Aberdeen AB21 9SB
Scotland
Email: claus@bioss.sari.ac.uk

Marc Morant
The Royal Veterinary and Agricultural University
Department of Plant Biology
Thorvaldsensvej 40, opgang 6, sal 2
DK-1871 Frederiksberg C, Denmark
Email: marmo@kvl.dk

Tobias Mourier
Department of Evolutionary Biology
University of Copenhagen
Universitetsparken 15
2100 Copenhagen Ø, Denmark
Email: tmourier@zi.ku.dk

Kim Nielsen
Molecular Diagnostic Laboratory
Aarhus University Hospital
Skejby
DK-8200 Aarhus N, Denmark
Email: kini@kba.sks.au.dk

Pernille Nielsen
Bioinformatik Centret
University of Copenhagen
Universitetsparken 15
2100 Copenhagen Ø, Denmark
Email: pern@binf.ku.dk

Ståle Nygaard
Institutt for Eksperimentell Medisinsk Forskning
Ullevål Universitetssykehus
Kirkeveien 166
N-0407 Oslo, Norway
Email: stale.nygard@basalmed.uio.no

Louise Olofsson
University of Gothenburg
Vita Straaket 12, pav 8:3
SE-41345 Göteborg
Sweden
Email: Louise.Olofsson@medic.gu.se

Yudi Pawitan
Department of Medical Epidemiology
Karolinska Institutet
P.O. Box 281
SE-171 77 Stockholm, Sweden
Email: Email: Yudi.Pawitan@mep.ki.se

Jose M. Pena
Decision Support Systems, Dept. of Computer Science
Aalborg University
Fredrik Bajers vej 7E
9220 Aalborg, Denmark
Email: jmp@cs.auc.dk

Camilla Büchler Seier Petersen
Novozymes A/S
Smørmosevej 25
1B1.12
DK-2880 Bagsværd, Denmark
Email: cbsp@novozymes.com

Jørgen Petersen
Department of Biostatistics
Panum Instituttet
Blegdamsvej 3
2200 Copenhagen N, Denmark
Email: j.h.petersen@biostat.ku.dk

Thomas Agersten Poulsen
Novozymes A/S
Smørmosevej 25
2CS.44
DK-2880 Bagsværd, Denmark
Email: tapo@novozymes.com

Mark Reimers
Center for Genomics and Bioinformatics
Karolinska Institutet
Berzelius Väg 35
SE-171 77 Stockholm, Sweden
Email: mark.reimers@cgb.ki.se

Mats Rudemo
Mathematical Statistics
Chalmers University of Technology
S-412 96 Göteborg
Sweden
Email: rudemo@math.chalmers.se

Chiara Romualdi
CRIBI Biotechnology Center University of Padova
Via U. Bassi 58/B
35100 Padova
Italy
Email: chiara@cribi.unipd.it

Jesper Ryge
Dept. of Neuroscience, Kiehn Lab.
Karolinska Institutet
Retziusväg 8
SE-17177 Stockholm, Sweden
Email: jesper.ryge@neuro.ki.se

Ursula Sauer
ARC Seibersdorf Research
Biotechnology
A-2444 Seibersdorf
Austria
Email: Ursula.Sauer@arcs.ac.at

Mikko Sillanpää
Rolf Nevanlinna Institute
University of Helsinki
P.O.Box 4
FIN-00014 University of Helsinki, Finland
Email: mjs@rni.helsinki.fi

Anders Sjögren
Department of Mathematical Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Email: anders.sjogren@math.chalmers.se

Lars Snipen
Department of Mathematical Sciences
P.O. Box 5035
N-1432 Aas
Norway
Email: lars.snipen@imf.nlh.no

Christine Steinhoff
Max Planck Institute for Molecular Genetics
Dept. Computational Molecular Biology
Ihnestr 73
14195 Berlin, Germany
Email: steinhof@molgen.mpg.de

Anna Svensson
Department of Clinical Physiology
Karolinska Institutet at Huddinge University Hospital
SE-11666 Stockholm
Sweden
Email: anna.svensson.9056@student.uu.se

Michael Sørensen
Department of Applied Mathematics and Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø, Denmark
Email: michael@math.ku.dk

Peter Sørensen
Department of Animal Breeding and Genetics
Danish Institute of Agricultural Sciences
Postbox 50
DK-8830 Tjele, Denmark
Email: pesorens@vt.edu

Qihua Tan
Odense University Hospital
Sdr. Boulevard 29
5000 Odense C
Denmark
Email: qihua.tan@ouh.fyns-amt.dk

Thomas Thykjaer
Department of Clinical Biochemistry
Aarhus University Hospital
Skejby
DK-8200 Aarhus N, Denmark
Email: thykjaer@kba.sks.au.dk

Mark van de Wiel
Eindhoven University of Technology
TUE, Department of Mathematics and Computer Science
P.O. Box 513
5600 MB Eindhoven, The Netherlands
Email: markvdw@win.tue.nl

Marjan van Erk
Wageningen University
Tuinlaan 5
6703 HE Wageningen
The Netherlands
Email: marjan.vanerk@wur.nl

Paolo Vidoni
Department of Statistics
University of Udine
Via Treppo 18
I-33100 Udine, Italy
Email: vidoni@dss.uniud.it

Mark van der Laan
Division of Biostatistics
University of California
140 Earl Warren Hall
Berkeley, CA 94720-7360, USA
Email: laan@stat.berkeley.edu

Anja von Heydebreck
Max-Planck-Institute for Molecular Genetics
Department of Computational Molecular Biology
D-14195 Berlin
Germany
Email: anja.heydebreck@molgen.mpg.de

Michael Væth
Department of Biostatistics
University of Aarhus
Vennelyst Boulevard 6
DK-8000 Aarhus C, Denmark
Email: vaeth@biostat.au.dk

Stefan Winter
Mathematisches Institut
Universität Stuttgart
Nansenstr. 22
D-71522 Backnang, Germany
Email: sh.winter@web.de

Ernst Wit
Department of Statistics
University of Glasgow
15, University Gardens
Glasgow G12 8QW, United Kingdom
Email: ernst@stats.gla.ac.uk

Torben F. Ørntoft
Department of Clinical Biochemistry
Aarhus University Hospital, Skejby
DK-8200 Aarhus N
Denmark
Email: orntoft@kba.sks.au.dk