Critical Assessment of Microarray Data Analysis

Vienna, December 4th-6th

PROCEEDINGS

TECAN. Agilent Technologies Vienna

WHAT IS CAMDA?

CAMDA (1,2) was founded to provide a forum to critically assess different techniques used in microarray data analysis. It aims to establish the state-of-the-art in microarray data analysis, as well as identify progress and highlight promising directions for future efforts. In order to achieve these goals, CAMDA adopted the approach of community-wide experiment, letting the scientific community analyse the same contest data sets. Researchers worldwide are invited to take the CAMDA challenge. Accepted contributinos are presented in short talks (25 mins), and the results of analysis are discussed and compared at the CAMDA conference. Posters provide an additional opportunity of presenting and discussing work. As a special opportunity, this year, a selection of analysis predictions will also be verified experimentally by the laboratory collecting the original contest data set.

CAMDA, which began in 2000, was initiated by Simon Lin and Kimberly Johnson from the Duke University Bioinformatics Shared Resource. It is patterned after the molecular modeling community's well-known CASP (3) experiment. In this sense, CAMDA is a functional genomics successor of the other well-known community-wide experiments, such as GASP (4) in genomics, CASP (3) in protein modeling, GAW (5) in statistical genetics, and PTC (6) in computational toxicology.

The first CAMDA conference (CAMDA'00) was held December 18–19, 2000. Attended by 250 biologists, statisticians, computer scientists and mathematicians from 7 countries, the conference truly brought together the major players in this field. Since then the CAMDA conference has grown stronger and more exciting, and has regularly been featured in top journals such as the Nature (1,2) and a recent Nature Methods editorial (7). Come join this exciting conference!

In 2006 it was decided that CAMDA would become a roving conference. This new period begins with Camda 2007, organised by Joaquin Dopazo at the CIPF in Spain, the first stop in the forecoming years of international roving. This year, Boku University will host CAMDA in Vienna, Austria.

All individuals and groups from both academic and commercial entities are invited to join the award competition.

1. Johnson, K.F. and Lin, S.M. (2001). Call to work together on microarray data analysis. Nature 411, 885.

2. Tilstone, C. (2003) Vital Statistics. Nature 424, 610

3. CASP: Critical Assessment of Techniques for Protein Structure Prediction. http://predictioncenter.llnl.gov

4. GASP: Genome Annotation Assessment Project. http://www.fruitfly.org/GASP1

5. GAW: Genetic Analysis Workshop. http://www.sfbr.org/gaw/

6. C. Helma, R. D. King, S. Kramer, and A. Srinivasan (2001). The Predictive Toxicology Challenge 2000–-2001. Bioinformatics 17, 107.

7. (2008) Going for algorithm gold, Nature Methods 5, 569. (link to journal)

CONTEST DATASETS

Primary dataset

The laboratories of Prof. Cristin Print and collaborators make available raw and processed data from a small microarray gene expression time-course experiment that is typical of gene expression time-course data sets yet provides an unusual opportunity for pushing the performance of analysis methods.

The experiment recorded the response of human vascular endothelial cells to serum withdrawal, triggering apoptosis. Apoptosis is known to be a major process for tissue remodelling during development and homeostasis in the adult, and also has a central role in many diseases. An initial, preliminary analysis and discussion of this data set has been published (1) and provides a good introduction to the biological background and context of the experiment.

The data set is typical in that a complex biological phenomenon is probed by a timecourse with only a few measurements, in this case 8 time-points and 3 replicate pools of cells from 10 distinct individuals each. It thus provides the *classical challenge of microarray data analysis of extracting insight in a data space of very uneven dimensionalities*, in this case 20k variables x (8×3) measurements. Also, a large number of independent experiments and established knowledge is available regarding apoptosis. Taking advantage of such external information for inference is again a typical challenge of the field.

The experiment is *unusual*, however, in that by design its focus is on detecting possible early causes. The challenge hence is rather to ***identify candidate regulators*** rather than primarily their targets by concentrating on very early time-points. We believe that this type of challenge will become a more and more central task for microarray analysis, particularly when considering the platform's strength in *detecting low-copy-number molecules, such as transcription factors, that potentially drive later transcriptional events*. The development of improved algorithms in this area will therefore continue to grow in relevance. The performance of new algorithms, however, is hard to assess without additional laboratory experiments. As part of every year's analysis challenge, the Program Committee will vote for the most interesting analysis. For this contest, Prof. Print's laboratory has kindly offered to ***experimentally test predictions*** *by siRNA knock-down of the most promising candidates emerging* (with a budget for costs of 5–10kNZ$). Together with the experimental design this offers a *special opportunity for developing and testing novel integrative algorithms for the detection of regulatory factors from typical (small) time-course data sets and available external knowledge*.

(1) Affara M, Dunmore B, Savoie C, Imoto S, Tamada Y, Araki H, Charnock-Jones D.S, Miyano S, Print C. (2007) 'Understanding endothelial cell apoptosis: what can the transcriptome glycome and proteome reveal?' *Phil. Trans. R. Soc. B.* **362**, 1469–-87.

Emerald dataset : A Microarray Experiment to Study the Relative Magnitudes of Technical and Biological Variation

Microarray science and technology has progressed to the point at which careful work yields reliable measurements. There is a growing understanding of the sources of variability in microarray experiments, and ways to control that variability are propagating. In part because the technical variability observed in contemporary microarray experiments has become better controlled, statistically significant lab-to-lab and batch-to-batch effects have been observed. A number of experiments which study the same samples across a variety of laboratories and platforms have reported this. The essential question is whether these effects are significant with respect to the biological variability observed amongst the samples. This question lies at the heart of establishing the fitness for purpose of microarrays for biological studies.

We have data available produced by *three different laboratories measuring the same samples on three different platforms* – each with their own *batch factors* (Liggett *et al.*, 2008). The platforms are the *Affymetrix Rat Genome 230 2.0* array, the *Illumina RatRef-12* array, and the *Agilent Whole Rat Genome* array. The samples are a *titration mixture* of RNA isolated from kidney and liver, from 6 different normal control rats from an earlier experiment at Novartis. This titration presents a series of 4 samples from each rat: RNA from the kidney, a mixture of 75% RNA from kidney and 25% from liver, a mixture of 25% RNA from kidney and 75% from liver, and RNA from the liver. These samples were measured in replicate, for each animal. Pooled samples from the various animals were also measured, for a nominal 96 arrays from each platform.

The relationship amongst these samples enables model-based analysis, amongst other approaches. Model-based approaches can be compelling because they permit observation and apportionment of variation in the residuals. The titration samples present interesting opportunities for alternative analyses as well, with the titration fraction as a surrogate or proxy for RNA concentration.

A particular interest for this CAMDA dataset is its use for evaluating the performance of different preprocessing approaches and techniques. We encourage research groups to address this question. Assessment using a model-based approach might enable estimation of any bias that might be introduced in preprocessing. Such estimates would, for the first time, provide valuable quantitative insight to enable the microarray data analysis community to make appropriate compromises when selecting a preprocessing pipeline.

(2) Ligget W, Peterson R, Salit M. (2008) 'Technical vis-à-vis biological variation in gene expression measurements', preprint.

INFORMATION

Location

The conference is held at the Muthgasse institutes of Boku University Vienna, in fast and easy reach by public transport (e.g., tube U4 Heiligenstadt) or car from the city. The institute address is AT-1190 Muthgasse 18, and meetings will be held in the lecture room XXI on the ground floor of the main buildings.

See and do

The city of Vienna is renowned for its beauty, culture, and high general quality of life, so do plan a longer stop if you can take some time out! A lot of information is provided in the information pack in your bag or online at http://www.wien.gv.at/english/.

There will be a social programme featuring highlights such as a cocktail reception at the world-famous City Hall including live music and generous buffet dinner in the Senate Chamber. We will also organize a visit to a Heuriger style pub.

We wish you a lovely stay in Vienna and hope you can take the time to explore the city and the stunning countryside that is in easy reach!

Public transport / tickets

Vienna has an excellent public transport system. Many hotels sell tickets at reception, else they are available from machines in all tube stations and on many tram stations. These accept banknotes and credit cards. You will need to get a ticket before boarding a train. While you can buy tickets in trams and busses, they are more expensive than when you get a ticket before boarding.

Perhaps of interest, if you are interested in sight-seeing and museums, for €18.50, there is the `Vienna card´ for tourists, which entitles you to 72h of unrestricted public transport and serves as a discount pass to many museums, as well as several other attractions and restaurants.

If you are staying longer or are travelling in company, then the `8 day-tickets´ carnet is interesting. Note that this is not an “8-day ticket” but rather allows you to use up 8 individual `day-tickets´ as you go, and even share the carnet amongst people. So, if you are travelling with a partner, say, you could insert two segments of the ticket into the validation machine, which will entitle two people to unrestricted travel for the day. It costs €27.20 and at €3.40 per day is the cheapest option as soon as you make at least two trips per day (which you almost certainly will).

Lastly, there are 72h/48h/24h and single-trip tickets priced at €13.60/€10.00/€5.70 and €1.70.

PROGRAM

Thursday, 4 Dec.

16.00 - 17.00 *Registration*

17.00 - 17.15 Welcome address

17.15 - 18.15 **Keynote**, 'Towards cracking the code of transcription and chromatin regulation', Eran Segal, Weizmann Institute, Israel.

18.15 - 18.45 'Analysis of comparative genomic hybridization and SNP arrays for the detection of chromosomal aberrations in single cells', Peter Konings**,** *et al.*, Katholieke Universiteit Leuven, Belgium.

20.00 - late *Cocktail reception* / *dinner*, City Hall (Rathaus), Senate Chamber

Friday, 5 Dec. – Emerald sessions

09.00 - 09.30 Introduction to the Emerald dataset, Ron Peterson, Novartis Institute of Biomedical Research, Cambridge, Massachusetts, U.S.A.

09:30 - 10:15 **Keynote**, 'Muddling of modelling your way through normalization?', Ernst Wit, University of Groningen, The Netherlands.

10.15 - 10.35 *Coffee* / Poster session

10.35 - 11.20 'Metrology for Gene Expression: Measurement Batch Effects, Probe Sensitivity, Gene-List Reproducibility', Walter Liggett, NIST, Gaithersburg, Maryland, U.S.A.

11:20 - 11:50 'Intrinsic metrics for hybridization control and global expression profiling – the fruit fly developmental time series', Hans Binder, Mario Fasold, and Jan Brücker, IZBI, Universität Leipzig, Germany.

11:50 - 12:10 'Experiment quality – direct route to reliable data', Ralph Beneke, Tecan Austria.

12:10 - 13:30 *Lunch*

13:30 - 14:15 **Keynote**, 'An array of FDA efforts in pharmacogenomics', Weida Tong, MAQC consortium, Toxicoinformatics, FDA, Jefferson, Arizona, U.S.A.

14:15 - 14:45 'Identification of spatial biases in Affymetrix oligonucleotide microarrays', Jose Manuel Arteaga-Salas, *et al.*, BEAMS, University of Essex, Colchester, U.K.

14:45 - 15:15 'EMERALD microarray platform comparison based on hypothesis tests under order restrictions', Florian Klinglmüller and Thomas Tuechler, Department of Statistics and Probability Theory, University of Technology Vienna, Austria.

15:15 - 15:45 *Coffee* / Poster session

15:45 - 16:15 'Exploiting the EMERALD mixture design for model based microarray platform comparisons by Bayesian inference of technical and biological variance components'. Thomas Tuechler, *et al.*, Chair of Bioinformatics, Boku University Vienna, Austria.

16:15 - 16:45 'Progress on transformation and normalization ontology', James Malone, European Bioinformatics Institute (EMBL-EBI), Cambridge, U.K.

16.45 -17.15 Panel discussion

19.00 - late Dinner in a typical local 'heuriger' style restaurant in town (optional)

Saturday, 6 Dec.

09.00 - 09.45 **Keynote**, 'Multiple Testing on the Graph of Gene Ontology', Jelle Goeman, Leiden University Medical Center, The Netherlands.

09.45 - 10.15 'Effect of Single Nucleotide Polymorphism (SNP) in Affymetrix probes', Olivia Sanchez-Graillet, William B. Langdon, and Andrew Harrison, BEAMS, University of Essex, Colchester, U.K.

10.15 - 10.45 'Extending pathways with inferred regulatory interactions from microarray data and protein domain signatures', Tim Beissbarth, Molecular Genome Analysis, German Cancer Research Center (DKFZ), Heidelberg, Germany.

10.45 - 11.15 *Coffee* / Poster session

11.15 - 11.45 'Modeling of microarray time-course data with dynamic Bayesian networks and within-time-point interaction', Brian Godsey and Peter Sykacek. Chair of Bioinformatics, Boku University Vienna, Austria.

11.45 - 12.15 'Inference of Key Transcriptional Regulators in Endothelial Cell Apoptosis using Bayesian State Space Models', David Wild, Claudia Rangel-Escareno, and Irma Aguilar-Delfin, Systems Biology Centre, University of Warwick, U.K.

12.15 - 12.30 Closing words

There social programme includes a cocktail reception at the world-famous new-gothic City Hall, complete with a generous buffet dinner in the Senate Chamber and live piano music. We will also arrange an optional trip to a local Heuriger country style restaurant/pub.

TALKS

Keynote

Towards cracking the code of transcription and chromatin regulation

Eran Segal, Weizmann Institute, Israel

The detailed positions of nucleosomes profoundly impact gene regulation and are partly encoded by the genomic DNA sequence. However, less is known about the functional consequences of this encoding. We address this question using a genome-wide map of millions of yeast nucleosomes that we sequenced. Utilizing the high resolution of our map, we refine our understanding of how nucleosome organizations are encoded by the DNA sequence, and demonstrate that the genomic sequence is highly predictive of the in vivo nucleosome organization, even across new nucleosome-bound sequences that we isolated from fly and human. We find that Poly(dA:dT) tracts are an important component of these nucleosome positioning signals, and that their nucleosome-disfavoring action results in large nucleosome-depletion over them and over their flanking regions, and enhances the accessibility of transcription factors to their cognate sites. Our results suggest that the yeast genome may utilize these nucleosome positioning signals to regulate gene expression with different transcriptional noise and activation kinetics, and DNA replication with different origin efficiency. These distinct functions may be achieved by encoding both relatively closed (nucleosome-covered) chromatin organizations over some factor binding sites, where factors must compete with nucleosomes for DNA access, and relatively open (nucleosome-depleted) organizations over other factor sites, where factors bind without competition.

Analysis of comparative genomic hybridization and SNP arrays for the detection of chromosomal aberrations in single cells

Peter Konings[1], Evelyne Vanneste[2,3,7], Thierry Voet[2,7], Cédric Le Caignec[2,*], Michèle Ampe[4], Cindy Melotte[2], Sophie Debrock[3], Mustapha Amyere[5], Miikka Vikkula[5], Frans Schuit[6], Jean-Pierre Fryns[2], Geert Verbeke[4], Thomas D'Hooghe[3], Joris R Vermeesch[2] & Yves Moreau[1]

[1]ESAT-SISTA, K.U.Leuven, Leuven, Belgium.

[2]Center for Human Genetics (CME), K.U.Leuven, Leuven, Belgium.

[3]Leuven University Fertility Center (LUFC), University Hospital Gasthuisberg, Leuven, Belgium.

[4]Biostatistical Center, K.U.Leuven, Leuven, Belgium.

[5]de Duve Institute, Université Catholique de Louvain, Brussels, Belgium.

[6]Molecular Cell Biology, Gene Expression Group,University Hospital Gasthuisberg, Leuven, Belgium.

*Present addresses: Service de Génétique Medicale, Centre Hospitalier Universitaire, Nantes, France; INSERM, U915, Nantes, France; Université de Nantes, Faculté de Médecine, l'Institut du Thorax, Nantes, France

Correspondence should be addressed to Peter Konings and Yves Moreau, ESAT-SISTA, K.U. Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.

Genomic arrays, including Comparative Genomic arrays and SNP arrays, enable the detection of genetic variation among individuals or among cell populations. Thanks to the high reproducibility of measurements of DNA levels (by contrast to RNA levels), genomic arrays have emerged as a highly effective application of microarray technology in oncology and in genetics of both congenital and complex disorders. A major emerging challenge for genomic arrays is genotyping and copy number determination using genomic content from a SINGLE cell. This capability is of interest in several situations where mosaicism (genetic variability at the cellular level) is present, such as in certain types of cancer and in embryonic development.

Recently, we have developed experimental methods to amplify single-cell DNA content and present here analysis strategies to determine the copy number status of single cells in early embryos, using both in-house arrays of genomic clones (BAC) and Affymetrix 250K GeneChip SNP arrays. One of the major challenges is the amount of noise and bias that is introduced by the amplification step. A mixture model was developed to detect copy number variation using the BAC array, while correcting for amplification bias. Results from the BAC array were combined with those from the CNAG and CNAT tools on the SNP array, showing a high concordance. Astonishingly, our analysis revealed chromosomal imbalances in about 90% of embryos obtained from couples with normal fertility. The observation of such a high level of genomic instability in early embryos has significant implications for our understanding of embryonic development and for clinical applications of genetic screening.

Keynote

Muddling or modelling your way through normalization?

Ernst Wit, University of Groningen, Netherlands

Preprocessing is typically thought of a "data cleaning" activity, separate from further inference. There are obvious advantages to this view: this way, the analyst does can apply an array of methods to the "cleaned" data, without every time having to worry about possible artifacts. However, for some standard tasks, such as detecting differential expression, a joint model belongs to the possibilities. Moreover, the particular structure of the CAMDA dataset this year lends itself particularly to modelling different aspects via a mixed effects model.

Metrology for Gene Expression: Measurement Batch Effects, Probe Sensitivity, Gene-List Reproducibility

Walter Liggett, National Institute of Standards and Technology
Jean Lozach, Illumina
Anne Bergstrom Lucas, Agilent
Ron Peterson
Marc Salit, National Institute of Standards and Technology
Danielle Thierry-Mieg, National Center for Biotechnology Information, NIH
Jean Thierry-Mieg, National Center for Biotechnology Information, NIH
Russ Wolfinger, SAS

As for other measurements, metrology for gene expression involves issues such as sources of measurement variation, measurement calibration, and inference on comparisons. Insight into these issues in the highly multiplexed case of gene expression measurement is possible on the basis of the EMERALD dataset of CAMDA08. This dataset contains measurements on the RNA of six animals (*rattus norvegicus)* made with Affymetrix, Agilent, and Illumina platforms. For each animal, there are replicate measurements on the liver RNA, the kidney RNA, and mixtures of these two RNAs.

We have obtained insight into the relative size of measurement batch effects and biological variation as represented by the animal-to-animal differences. These differences provide a practical benchmark because the animals were all subject to the same control-group treatment.

Although calibration curves for individual probes are unknown, insight into calibration can be obtained from a platform-to-platform correspondence that identifies probes that measure the same transcript. This identification allows insight into the relative sensitivity of probes from different platforms.

For biologists, gene expression microarrays provide an approach to identifying genes with particular properties such as change in expression with experimental treatment. The genes thus identified populate a gene list. In simple case-control studies, there are individuals in two groups that are each subject to a different experimental treatment. Because each group separately exhibits biological variation, the identification criterion usually

involves a statistical test of the null hypothesis that the difference between the groups is solely the result of this biological variation. A gene list is obtained by applying this statistical test to the measurement set for each gene. Because we have measurements on six animals, we can obtain insight into such gene lists.

We have obtained some general observations on the metrology issues considered. First, although the animal-to-animal variation is generally larger than the measurement batch effects, our measurements do lead to the conclusion that these effects should not be ignored in experimental design and analysis. It is moreover the case that the measurement batch effects might be larger in a different experiment. Second, over the set of transcripts for which liver expression is appreciably different from kidney expression, no platform is undeniably more sensitive than another. However, the difference in probe sensitivity between two platforms varies appreciably from transcript-to-transcript. That is, one platform seems more sensitive for some transcripts, and the other platform more sensitive for other transcripts. This observation suggests considerations in the interpretation of single-platform studies. Third, we find that gene list reproducibility is likely to be worse than might be expected.

The experiment described here offers an approach to measurement system insight that could feasibly be part of any substantial gene expression study. There are reasons why one might want to change the design of our experiment. Inclusion of more animals would lead to more insight into gene-list reproducibility. Our investigation provides full coverage only of the probes for which liver expression differs from kidney expression. Inclusion of more animal organs would lead to better coverage of the probes.

Intrinsic metrics for hybridization control and global expression profiling – the fruit fly developmental time series

Hans Binder*, Carolin Ulbricht, Mario Fasold and Jan Brücker

University of Leipzig, Interdisciplinary Centre for Bioinformatics; * corresponding author: binder@izbi.uni-leipzig.de

Abstract

Quality control and calibration of microarray data account for detection and correction of technological variation. We present a new calibration approach which generates a chip-specific metrics using the intensity data of each particular GeneChip. This so-called hook method assesses the performance of a given hybridization based on a set of chip-related summary characteristics. The method is applied to a developmental time-series to study the effect of technological and biological factors on the variability of the data and to characterize global expression changes.

Introduction

The process of producing microarray data involves multiple steps which may suffer from different error sources resulting in poor expression data. Quality control is therefore an essential prerequisite for downstream expression analysis. In some cases the detection of data of "poor" quality however might be a misinterpretation of biological variation as technologically caused errors. Especially developmental and intervention experiments are candidates for global and/or unbalanced changes of the expression level which may pretend data quality problems.

The quality of pre-processed fruit fly data were recently assessed using different numerical measures such as the normalized unscaled standard error (NUSE) and the relative log-expression (RLE) the interpretation of which turned out to be problematic in the case of time-course data [1,2]. Particularly, NUSE estimates the variability of expression measures and RLE their deviation from the mean over the considered series after RMA-preprocessing including quantile normalization. Their use for quality assessment is based on two assumptions: (i) the majority of probed genes remain biologically invariant and (ii) up- and down-regulations compensate each other. In fact, developmental microarray data potentially violate both assumptions because the fraction of differentially expressed genes might be relatively high and inhomogeneous over time. Another problem arises from the multichip character of RMA-preprocessing which makes model fitting problematic for small numbers of replicates as typically collected in time course experiments.

In this paper we reanalyzed the fruit fly developmental series [3] using a novel method of microarray data calibration. This so-called hook method is a single-chip approach which independently processes the raw intensities of each microarray. It generates a series of chip characteristics suited for quality control and the assessment of the global expression degree. We illustrate the performance of the method and discuss its potency in the context of quality control and the analysis of large-scale unbalanced expression changes. We chose the fruit fly time series as an exemplary example because it allows direct comparison with the results of recent studies addressing similar issues [1,2]. In addition we generated a quality report of this series using several established quality measures provided by BioConductor R-routines [4] as supplementary information available via www [5].

Hook method

The so-called hook method (see [6-8] for a detailed description) applies to microarrays of the GeneChip-type containing pairs of perfect match (PM) and mismatch (MM) probes to estimate the abundance of ten thousands transcripts in one measurement. It independently analyzes the intensity data of each GeneChip microarray using the two-species Langmuir hybridization isotherm which assumes competitive binding of specific and "representative" non-specific transcripts to each probe. The method processes the PM and MM probe intensities (I^{PM} and I^{MM}, respectively) using the transformation

$$\Delta = \log I^{PM} - \log I^{MM} \quad \text{and} \quad \Sigma = \tfrac{1}{2}\left\langle \log I^{PM} + \log I^{MM} \right\rangle_{set} \quad , \qquad (1)$$

where $\langle\ldots\rangle_{set}$ denotes averaging over each probe set of usually 11 PM/MM probe pairs addressing one transcript. Smoothing of the Δ-versus-Σ plot provides the hook curve which enables decomposition of the probe signals into contributions due to specific and non-specific hybridization by simple graphical analysis (see Figure 1) and subsequent correction of the intensities for sequence specific effects using the positional-dependent nearest neighbour model as standard (Figure 1). The corrected signals are re-plotted into Δ-versus-Σ coordinates and again smoothed to obtain the corrected version of the hook curve which allows identification of absent and present probes using a simple break criterion (Figure 1).

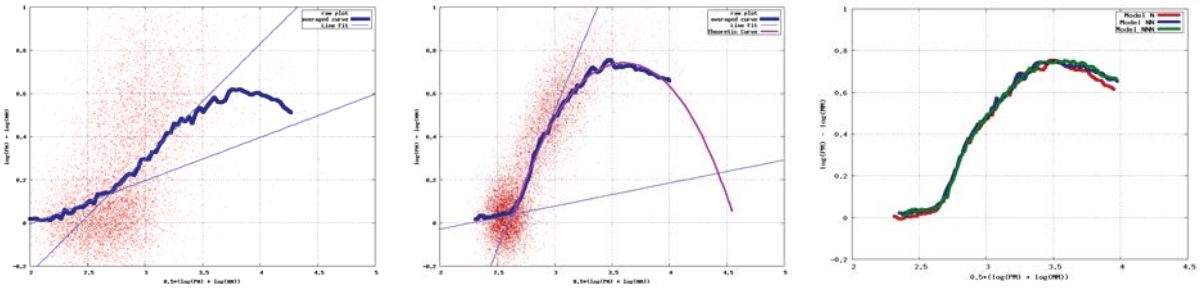


Figure 1: Hook-analysis of one Affymetrix Drosgenome 1 chip taken from the fruitfly time series: Raw (left panel) and corrected (middle panel) hook curves. Probe set level data are shown by scatter-dots. The intersection of the two lines defines the "breakpoint" which separates "absent" probe sets in the flat region of non-specific hybridization on the left from "present" probes in the increasing part of the curve on the right. The change of the slope indicates the onset of specific hybridization. The parabola-like curve in the middle panel shows the fit of the theoretical "hook-equation" (2) to the data. It provides four parameters of well-defined geometrical and physical meaning (see Figure 2). The right panel compares sequence-corrections using positional-dependent single-base (N), nearest neighbor (NN) and next nearest neighbor (NNN) models. Because of the negligible improvement NN→NNN we use the NN-model as standard.

Then, the hook curve is analyzed in terms of the two-species Langmuir binding model which predicts the following parametric equations for the Δ- and Σ-coordinates

$$\Delta(R) = \Delta^{start} + \log\left\{(R+1)/\left(R\cdot 10^{-\alpha}+1\right)\right\} - 1\log\left\{B^{PM}(R)/B^{MM}(R)\right\}$$

and

$$\Sigma(R) = \Sigma^{start} + \tfrac{1}{2}\log\left\{(R+1)\cdot\left(R\cdot 10^{-\alpha}+1\right)\right\} - \tfrac{1}{2}\log\left\{B^{PM}(R)\cdot B^{MM}(R)\right\} \quad . \qquad (2)$$

with the saturation terms $B^{PM}(R) = 1 + 10^{-\left(\beta - \frac{1}{2}\Delta^{start}\right)}(R+1)$ and $B^{MM}(R) = 1 + 10^{-\left(\beta + \frac{1}{2}\Delta^{start}\right)}\left(R\cdot 10^{-\alpha}+1\right)$.

The argument; the so-called S/N-ratio

$$R \equiv \left(K^{PM,S}/K^{PM,N}\right)\cdot\left([S]/[N]\right) \quad , \qquad (3)$$

is an expression measure related to the specific transcript concentration [S]. It is given in "intrinsic" units of the effective concentration of non-specific transcripts [N] and scaled by the ratio of the respective binding constants of the PM-probes. [N] and the binding constants are chip-specific values whereas [S] is specified for each transcript.

The two parameter couples (Σ^{start}, Δ^{start}) and (α, β) characterize the position and the geometrical dimensions of the hook curve in terms of the coordinates of their starting point and their width and height, respectively (Figure 2). They are related to well-defined hybridization characteristics of the selected chip:

$$\alpha = \log\frac{s}{n} \quad , \quad \beta = \tfrac{1}{2}\log n - \left\langle\log\left(K^{PM,N}\cdot[N]\right)\right\rangle_{chip} \quad , \quad \Delta^{start} = \log n \quad , \quad \Sigma^{start} = \log I_{max} - \beta \quad . \qquad (4)$$

Here, s and n are the "PM/MM"-gain parameters which are defined as the mean, chip-averaged ratios of the binding constants of the PM and MM probes for specific and non-specific hybridization, $s = \left\langle K^{PM,S}/K^{MM,S}\right\rangle_{chip}$ and $n = \left\langle K^{PM,N}/K^{MM,N}\right\rangle_{chip}$, respectively. I_{max} is the maximum intensity reached at complete saturation of the probes with bound transcripts.

The simple relation between the geometrical dimensions on the one hand and basic hybridization characteristics of the selected chip on the other allows the straightforward evaluation of the particular hybridization by visual inspection of the corresponding hook curve. For example, its width β simply reflects the mean level of non-specific hybridization and its height α-parameter estimates the PM/MM-gain due to the central mismatch of the MM. Part a – c of Figure 2 illustrate the effect of typical experimental factors such as changes of the optical settings and of the amount of RNA.

The position of a probe set along the hook curve characterizes its hybridization properties which are governed by the superposition of specific and non-specific binding and by saturation of the probe spots. Accordingly, one can divide the hook curve into five consecutive hybridization regimes (Figure 2). The argument of the hook-equation can be simply related to the difference of the hook coordinates of a selected probe set relative to the respective start values to a good approximation [7] as illustrated in Figure 2,

$$\log(R+1) \approx \left\{\left(\Sigma - \Sigma^{start}\right) + \tfrac{1}{2}\left(\Delta - \Delta^{start}\right)\right\} \quad . \qquad (5)$$

Hence, the hook curve spans a metrics system for expression estimates in intrinsic units which are defined by the

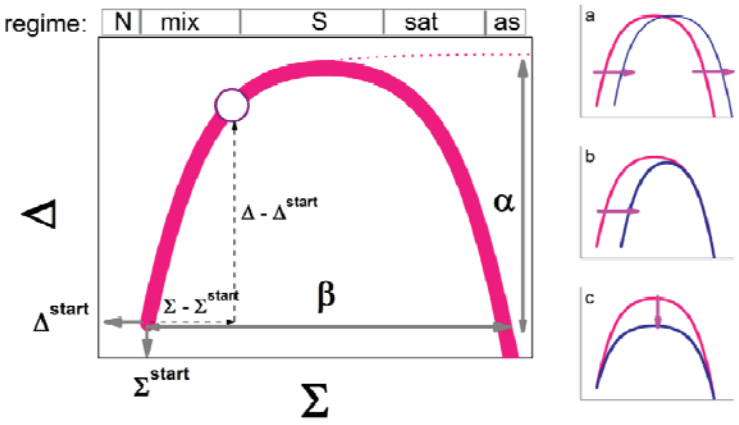


Figure 2: Geometrical dimensions of the hook curve (see Eqs. (2) and (4)) and hybridization regimes. The start coordinates (Σ^{start}, Δ^{start}) characterize the non-specific background level in intensity units and the N-PM/MM-gain, respectively. (β, α) characterize the width of the hook and its height in the absence of saturation, respectively. The curve divides into five hybridization regimes: non-specific (N), mixed, specific (S), partial (sat) and complete (as) saturation. The open circle represents the Σ,Δ-coordinates of a selected probe set (transcript). The distances relatively to the start point are directly related to the respective expression (Eq. (5)). Part a – c schematically illustrate typical experimental effects on the dimensions and/or position of the hook curve: a) Optical scaling of the intensity owing to changes of the scanner settings or the labeling equally shift the start AND end points in horizontal direction. b) Alterations of the non-specific background level owing to changes of the amount of RNA and/or of of its composition shift ONLY the start point in horizontal direction giving rise to the narrowing of the hook for larger background contributions. c) Modifications of the mismatch design change the vertical dimensions of the hook, e.g. a smaller PM/MM gain decreases its height.

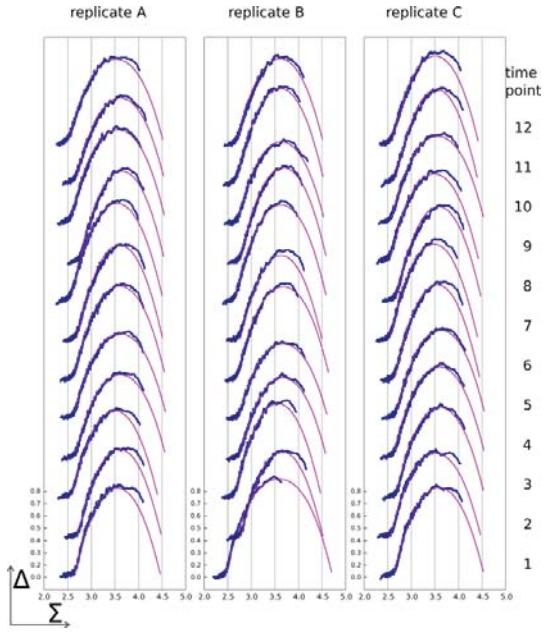


Figure 3: Corrected hook-plots of the fruit fly developmental time-series: RNA obtained from embryos of wild type fruit flies every hour during incubation over 12 hours was hybridized on Drosgenome 1 GeneChips. The time-series was performed in triplicate; A, B and C. The parabola-like curves are fits of Eq. (2).

level of non-specific hybridization and mean binding constants of a given chip (Eq. (3)). Averaging over all probe sets provides the S/N-exponent $\lambda = \left\langle \log\left(R+1\right)\right\rangle_{R>0\ 5;chip}$ as measure of the mean specific transcript abundance.

In the final step of the hook analysis the sequence-corrected probe level intensity data are corrected for the non-specific background and for saturation effects and then summarized for each probe set to get transcript-related expression estimates [7].

Hybridization control of the Drosophila time series

We reanalyzed the fruit fly developmental time series created by Tomancak et al. [3] (http://www.fruitfly.org/cgi-bin/ex/insitu.pl). It comprises three replicated series (A, B and C) of 12 consecutive time points starting at 1 hour and ending at 12 hours post egg laying of the flies.

Figure 3 shows the collection of corrected hook-plots generated from the chip data. Except of chip B-1 all plots reveal - on first sight - similar hybridization quality without evident outliers and marked mutual horizontal and/or vertical shifts of the curves. Detailed analysis basically confirms this impression for the mean non-specific background intensity (Σ^{start}), and the width- (β) and height- (α) parameters, which remain virtually constant over time (Figure 4). The experimental series is obviously characterized by virtually equal optical settings and hybridization conditions.

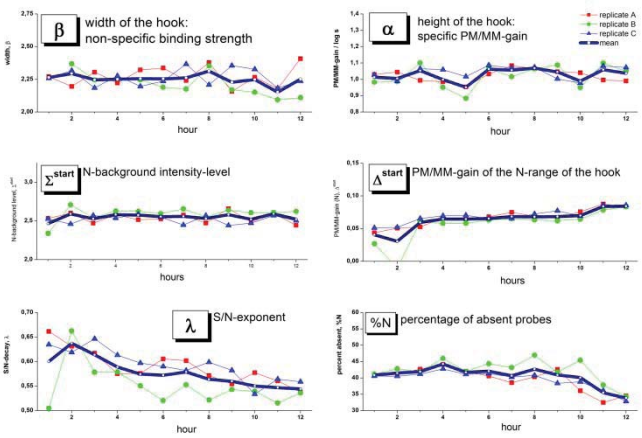


Figure 4: Selected hook-parameter as a function of fly-developmental time. The thin lines/symbols refer to the different replicates (A: red; B: green; C: blue) and the thick line to their mean. Series B systematically deviates from A and B (see β, λ and %N). It was hybridized on a different day than the other two series. This result confirms previous quality assessment using NUSE and RLE [2].

Contrarily, the PM/MM-gain of the N-range (Δ^{start}), partly the percentage of absent probes (%N) and especially the S/N-exponent (λ) indicate systematic trends upon development of the fly embryos which are probably not caused by technological effects. These changes thus possibly reflect subtle modifications of the global expression pattern which will be addressed in the next section.

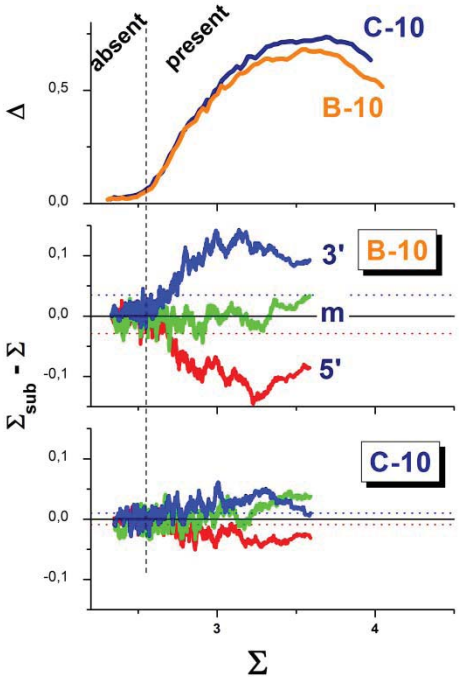


Figure 5: 3'/5'-amplification bias of transcript abundance. Corrected hook-plots (upper panel) and 3'/5'-difference plots of two selected chips (series B and C at hour 10) indicate different RNA-quality: The split between the 3'- and 5'-branches in the difference plot estimates the mean log-intensity difference between probe number 8-11 and 1-4 in each probe set ("m" refers to probes no. 5 – 7). It is plotted as a function of the total mean over all probes in each set Σ. Importantly, the 3'/5'-bias becomes evident only for "present" probes upon specific hybridization to the right from the breakpoint of the hook curves. In the N-range the 3'/5'-bias disappears. The total mean over the branches therefore markedly underestimates the 3'/5'-bias (see horizontal dotted lines). Sample B-10 is of poorer quality than sample C-10 because of the larger split in the range of present probes..

The amplification step upon RNA preparation potentially results in 5'-truncated transcripts with consequences for expression analysis. This amplification bias can be estimated using special control probe sets (e.g. GADPH) or the so-called RNA-degradation plot which log-averages the probe intensities according to their sequential ordering in the probe sets to identify poor RNA by relative large gradients along the probe rank [4].

We recently proposed a modification of this approach which accounts for the fact that non-specific hybridization is not affected by the 3'/5'-bias [8]. Our method calculates sub-averages Σ^{sub} over the first and the last four probes of each probe set (near to the 5' and 3' end of the transcript, respectively) using Eq. (1) and then the difference $\Sigma - \Sigma^{sub}$ is plotted as a function of the hook abscissa (Figure 5). A large split between the 3' and 5'-branches in the range of the present probes is indicative for poor RNA-quality. We identified only sample B-10 of poor 3'/5'-characteristics in the whole developmental series whereas the remaining chips are acceptable (see, e.g., C-10 in Figure 5). Note that the averaging over all probes without differentiation between specific and non-specific hybridization (as applied in the degradation plot) at best underestimates the 3'/5'-bias but at worst completely fails to detect poor RNA [8].

Global expression changes

To further explore the systematic trend revealed in Figure 4 we calculated the probability distribution of the S/N-ratio in terms of log(R+1) for early, intermediate and late stages of embryo development (Figure 6). The obtained plots can be roughly divided into three parts referring to absent (R=0), weakly-expressed (0< R < 2.2) and well-expressed (2.2 < R) genes. The abundance of the latter ones decays exponentially to a good approximation.

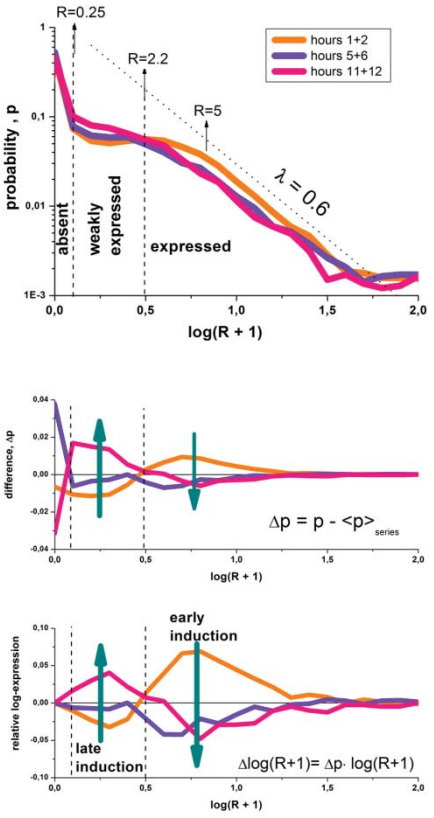


Figure 6: Global expression changes during fly-development. Probability distribution of the S/N-ratio, log(R+1), at early, middle and late stages: replicates A and C are pooled at time-points 1+2, 5+6 and 11+12, respectively. The distributions decay exponentially at larger abscissa values as illustrated by the dotted line (upper panel, λ is the decay constant). The two panels in the middle and below show the changes of the respective probability distributions relatively to their mean and the resulting effective change of the log-expression. Note that the different global changes of absent, weakly expressed and (moderate and strongly) expressed transcripts are induced more strongly at late or early stages of embryogenesis of the flies, respectively (see arrows).

The difference of the distributions with respect to their average reveals different trends in the tree expression ranges: The populations of well-expressed, of absent and of weakly expressed genes are maximal at early, intermediate and late stages of development, respectively. The global expression pattern thus gradually rearranges with time. The respective maximum net-changes of the expression degree (in terms of log(R+1)) are in the order of 5-10% of the expression values (lower panel of Figure 6).

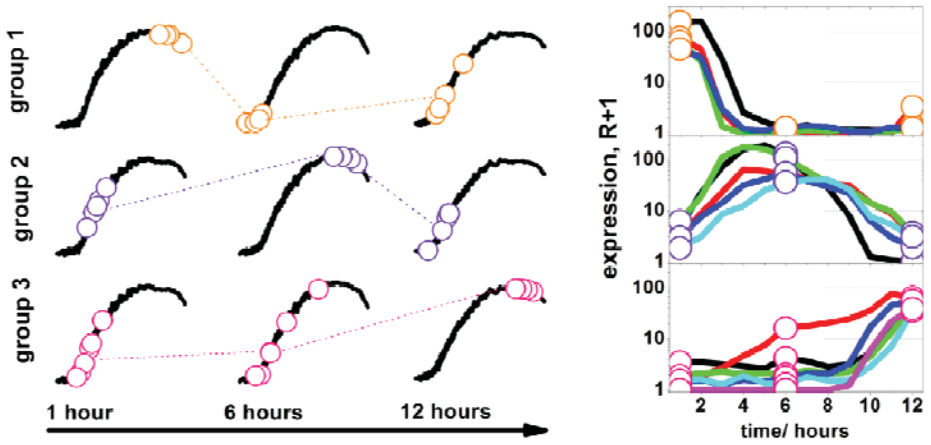


Figure 7: The hook-positions of selected genes reveal characteristic time-profiles: Group 1 refers to early induced, group 2 to broadly expressed and group 3 to late induced genes, respectively. The right part shows the full time profiles of the expression values of the selected genes. The open circles refer to the same points in both parts of the figure.

The observed global pattern of transcript abundance results from the superposition of very diverse expression changes of individual genes. Tomancak et al. analyzed in detail the temporal and spatial expression profiles of the fly embryos and relates them to different tissues and cellular functions [9]: About 35% of the studied 5560 genes show tissue-restricted expression with peaks especially at early or late stages of embryogenesis, 46% are broadly expressed showing a ubiquitous but not uniform profiles and for the remaining 19% no expression was detected. Figure 7 illustrates characteristic time courses of early induced, broadly expressed and late induced genes in terms of their hook coordinates (left panel) and expression values (right panel). The hook-coordinates in a simple and illustrative fashion reflect the expression degree in the context of their hybridization characteristics, for example, whether the probe signals are near the background or the saturation level or in the intermediate S- and mix-ranges (compare with Figure 2).

Conclusions

Changing variability of transcript abundance and unbalanced differential expression are inherent biological properties, which especially pronounce in developmental time-series experiments. Appropriate expression analysis requires subtle calibration techniques which correct raw data for technical artifacts without removing meaningful biological information from the corrected data. The hook-method provides expression estimates and chip-summary characteristics using the intrinsic metrics provided by the hybridization reaction. We demonstrate the potency of the method to establish new measures for microarray quality control which supplement existing standards of microarray quality assessment.

References

(1) Brettschneider, J.; Collin, F.; Bolstad, B. M.; Speed, T. P. *Technometrics* 2007, *MS06-032R*.

(2) Guan, S. H.; Zheng, J.; Brettschneider, J. *Proceedings of LASR'07* 2007, *(download: www.maths.leeds.ac.uk/lasr2007/proceedings/brettschneider.pdf)*.

(3) Tomancak, P.; Beaton, A.; Weiszmann, R.; Kwan, E.; Shu, S.; Lewis, S. E.; Richards, S.; Ashburner, M.; Hartenstein, V.; Celniker, S.; Rubin, G. *Genome Biol.* 2002, *3*, 0088.1.

(4) Gentleman, R. *AffyQC-report description* 2007.

(5) Rosolowski, M.; Binder, H. *report* 2008, *www.izbi.uni-leipzig.de/downloads_links/downloads/hook-QC_Affy_fruitfly.pdf*.

(6) Binder, H.; Preibisch, S.; Berger, H. *Methods in Molecular Medicine* 2008, *in press*, (preprint: www.izbi.de/izbi/working_papers.php).

(7) Binder, H.; Preibisch, S. *Algorithms for Molecular Biology* 2008, *3:12*.

(8) Binder, H.; Krohn, K.; Preibisch, S. *Algorithms for Molecular Biology* 2008, *3:11*.

(9) Tomancak, P.; Berman, B.; Beaton, A.; Weiszmann, R.; Kwan, E.; Hartenstein, V.; Celniker, S.; Rubin, G. *Genome Biol.* 2007, *8*, R145.

Experiment quality - direct route to reliable data

Ralph Beneke, Product Manager, Tecan Austria GmbH

Insufficient data quality? Wasting time, sample and money?
Start with a clear vision and explore and make use of existing collaborative networks & references. Plan your lab infrastructure & resources according to design of experiment (incl. quality control, optimization and validation process), the biological material and the workload & throughput.

Are there data reduction, -tracking and data management as well as appropriate budget & funding available for the routine?
Understanding principle of platforms – advantages & limitations of their analytical capability, robustness and flexibility is prerequisite when selecting appropriate platforms and tools.
The weakest step is the limiting step. Therefore measure of quality and variation of each step (success vs. failure rate) is crucial. Cost for maintaining the quality and gaining on flexibility to expand applications without losing on quality have to be considered.

No - or faulty - data is more expensive than generating accurate data.
Old saying “garbage in - garbage out” is true from design of experiment down to all steps until data analysis.
Learning from MAQC projects with good experimental design, optimization first, standardized processes and environmental conditions, applied analysis of variations on random block design and control measures. The robustness of platforms can buffer for smaller variations between different labs. Simplicity of workflow cuts labour cost, automation minimizes process complexity and failure rate (operator/protocol?) with the add-on of scalable throughput without losing on quality.

Tecan is more than delighted to listen to your different subjects and offer expertise in selecting from platforms and helping on implementing them in your lab routine.

Keynote

An Array of FDA Efforts in Pharmacogenomics

Weida Tong, MAQC consortium, Toxicoinformatics FDA, Jefferson, Arizona, U.S.A.

Pharmacogenomics (PGx) is identified in the FDA Critical Path document as a major opportunity for advancing medical product development and personalized medicine. An array of FDA efforts on PGx has been taking place at the inter-center and cross-community collaborative levels. Specifically, FDA issued guidance to industry on PGx data submission. The guidance defines a novel mechanism entitled Voluntary Genomics Data Submission (VGDS) whereby the sponsor is able to interact with the agency in the early stage of drug development by submitting PGx data on a voluntary basis. The name of Voluntary eXploratory Data Submission (VXDS) has been adopted recently to reflect diverse types of data received in this program, ranging from DNA microarray data, proteomics and metabolomics data to pharmacogenetic data including data from genome wide association study (GWAS) data. To facilitate this process, an integrated FDA bioinformatics tool, called ArrayTrack, developed at NCTR/FDA, is being refined as a review tool for managing, analyzing and interpreting this exploratory data i.e. genomic, proteomic and metabolomic data, from both clinical and non-clinical data submissions. To further understand the PGx technology in the regulatory context, a MicroArray Quality Control (MAQC) project has been initiated. This is a community supported project led by the FDA to address various issues associated with DNA microarrays, a critical technology used in the generation of PGx data. The lessens learned from both VXDS and MAQC are paving the way for development of the guidance document for future voluntary as well as regular submissions of PGx data to the FDA. ArrayTrack is an integral part of VXDS and MAQC. Together these capabilities provide the Agency with an integrated bioinformatics infrastructure to support data management, analysis and interpretation. ArrayTrack also serves as a vehicle to translate the guidance document into routine application for regulatory review and decision making.

Identification of spatial biases in Affymetrix oligonucleotide microarrays

J. M. Arteaga-Salas[1], G. J. G. Upton, W. B. Langdon and A. P. Harrison

Department of Mathematical Sciences, University of Essex, U.K.

Abstract

Affymetrix oligonucleotide microarrays contain spatial biases in their hybridizations. Some methods have been proposed to identify spatial biases in experiments with replicate arrays, however these methods are not useful when no replicates are available. In this work we propose a new method to deal with this problem.

1 Introduction

Affymetrix Oligonucleotide Microarrays have become a popular tool for analyzing gene expression measurements. In each array a number of probes (typically upwards of half a million) are arranged in a square grid where the genetic material is hybridized and measured in the form of light intensities. Each probe contains a sequence of 25 bases providing information about a designated gene. The design comprises pairs of "Perfect Match" probes (PM) and "Mismatch" probes (MM) that differ only in their central base; these are arranged in neighbouring locations across the array. Each probe belongs to a probeset (typically between 11-20 probes), and each probeset corresponds to a designated gene. All the probes in a probeset are arranged randomly across the array to avoid the presence of spatial biases. However, as we will show below, spatial biases in hybridization are not an uncommon problem in microarray experiments.

Several authors have reported methods to unmask spatial biases in microarray experiments. Suárez-Fariñas *et al.* (2005) developed the computational tool "Harshlight" for the automatic detection and masking of blemishes in microarrays. These are usually manifested as blobs of concentrated high/low values or as arcs and rings. Langdon *et al.* (2008) reported that these were most frequently seen towards the sides of an array, as are occasional "scratches".

Arteaga-Salas *et al.* (2008) showed that to identify spatial biases in an experiment with replicate arrays (biological or technical) it is useful to compare the values in each array with a reference value. The "Harshlight" package is also useful to identify biases in this type of experiments. However, these methods are useless if the experimental design does not include replication. In the next sections we will propose an alternative to identify spatial biases when no replicates are available.

[1]Contact: jmarte@essex.ac.uk. The work of J.M.A.S. is supported by CONACYT (Mexico) grant number 178596.

2 Identification of spatial biases

To identify spatial biases in experiments with three or more replicate arrays (biological or technical) Arteaga-Salas *et al.* (2008) suggested comparing the values in each location of the array with a reference value such as the median of the replicates at each location. By comparison with the reference value it is possible to identify locations with unusually high or unusually low intensity values. If such values are concentrated in regions of the array, then spatial biases can be visualized.

In the absence of replicates an alternative reference value is required. One alternative has been provided by Langdon *et al.* (2008) who calculated an "Average GeneChip" and a "Variance GeneChip" using all the Affymetrix Chips available[2] in the Gene Expression Omnibus (GEO) at http://www.ncbi.nlm.nih.gov/geo/ (separately by species and design). To obtain these arrays the geometric mean and the variance of the logarithm of the observed probe values in each available chip was calculated, discarding the upper and lower 0.5% of the values to avoid the effects of outliers. Using the average chip, A, and the variance chip, V, we can identify spatial biases in an observed array of logarithm values L of the same type following these steps:

1. For each location (i, j) in the array calculate the values h_{ij}, given by:

$$h_{ij} = \frac{L_{ij} - A_{ij}}{\sqrt{V_{ij}}}. \tag{1}$$

2. Sort the h_{ij} values by column j. For each sorted value assign a rank, and store them in array K.

3. To examine the spatial information in K define a sub-array centered at (i, j). The sub-array needs to include enough spatial information in a neighbourhood. We have verified that a sub-array size 11×11 is efficient for this purpose.

4. The sub-array centered at K_{ij} contains both PM and MM probes. Typically, the PM and MM intensities are highly correlated, therefore their ranks will be correlated too. To avoid using correlated values we do not consider physically adjacent cells. In other words we choose only one probe in a (PM,MM) probe pair (similar to a checkerboard pattern), thus selecting 61 locations from the total 121 in the 11×11 sub-array. Using the K-values for the selected cells we obtain the scores Z for each location, given by:

$$Z_{ij} = \frac{\sum_{n=1}^{61} K_n - 61 * \mu}{\sqrt{61 * \sigma^2}} \tag{2}$$

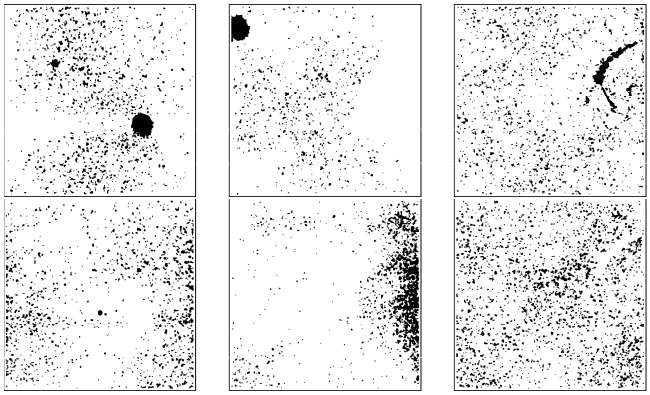
   where μ is the mean and σ^2 is the variance of a discrete uniform distribution (defined by the size of the chip). The scores Z will have a normal distribution with mean 0 and variance S^2 (in the absence of spatial biases $S^2 = 1$).

[2] As available in February 2007.

5. The locations where $abs(Z) \geq 2S$ have neighbourhoods with unusually high (or low) values. If these neighbourhoods are concentrated in the same region of the array, then a spatial bias is present.

To illustrate this procedure we selected three Affymetrix GeneChip Human Genome U133 Plus 2.0 non-replicate arrays from GEO (GSM46959, GSM76563 and GSM117700; all from accession number GSE2109). Figure 1 shows locations (i, j) where $abs(Z_{ij}) \geq 2S$. The three arrays show non-randomly concentrated regions for unusually high values (blobs and curves), and some concentrations for unusually low values in the edges of the array.

Figure 1: Spatial flaws in three HG-U133 Plus 2.0 arrays: (a) GSM46959 (left), (b) GSM76563 (center) and (c) GSM117700 (right). The upper row shows the locations of unusually large values and the lower row the locations of unusually small values.



3 Reduction of spatial biases

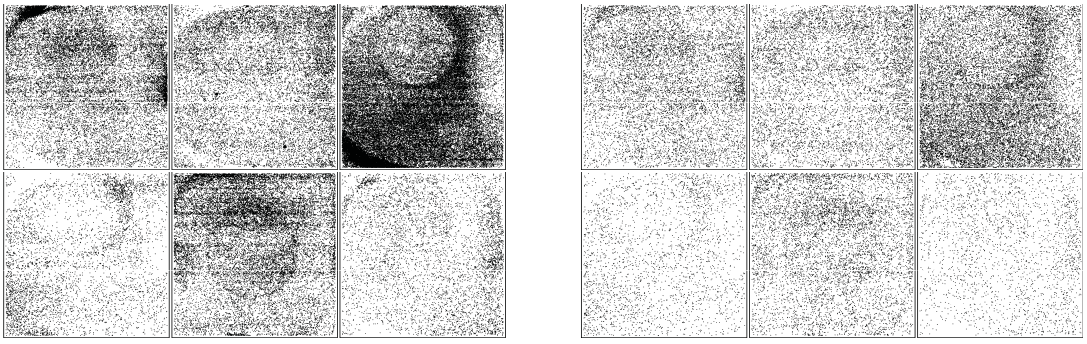
3.1 Comparison of arrays

To reduce the spatial biases found in Section 2 we need to compare each contaminated array with other arrays (at least two) of the same type from which any flaws have been removed. Arteaga-Salas *et al.* (2008) introduced two procedures to assist with flaw removal: complementary probe pair (CPP) and local probe effect (LPE). These methods can be used separately or in sequence. Typically, for replicate arrays the most effective method is to use LPE followed by CPP (denoted as LPE+CPP).

Figure 2 shows the results of using LPE+CPP for three replicates of the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array downloaded from GEO. The left hand side shows the spatial biases found in a set of three replicates (GSM96262-4; GEO accession number GSE4217) by plotting cells that differ by more than 25% of the median array value at each location, and the right hand side shows the few spatial biases that remain after correction with LPE+CPP.

We now compare two of these corrected arrays with each of the contaminated arrays in Section 2 to reduce the spatial biases. The CPP procedure is not appropriate in this context, so we use the LPE procedure which we now briefly describe.

Figure 2: (a) Spatial flaws in three replicates of the GSE4217 experiment (left) and (b) Remaining flaws in three replicates of the GSE4217 experiment after LPE+CPP (right). The upper row shows the locations of unusually large values and the lower row the locations of unusually small values.



3.2 Local probe effect (LPE) adjustment

This procedure can be used to reduce spatial biases whenever R arrays (which need not be replicates) are available. LPE uses the spatial structure in a 5×5 window centered at location (i, j) to decide whether adjustment should take place. For array r we first calculate the values d_{ijr} given by:

$$d_{ijr} = \frac{L_{ijr} - \alpha_{ij}}{\beta_{ij}} \tag{3}$$

where L_{ijr} is the logarithm of the observed value, α_{ij} is the median of the L_{ijr} values and β_{ij} is the standard deviation of the L_{ijr} values. We now define the quantities I_{ij} and G_{ij} as follows:

I_{ij} The identifier of the array corresponding to the case where d_{ijr} takes its largest absolute value.

G_{ij} This takes the value 1 if the d-value with the largest magnitude (at this location) was positive, and is otherwise equal to -1.

Using these two quantities, define the identifier E_{ij} by:

$$E_{ij} = I_{ij} \times G_{ij} \tag{4}$$

so that, with R arrays, E_{ij} takes one of the values $\{-R, -(R-1), ..., -2, -1, 1, 2, ..., (R-1), R\}$.

If the window contains a large number of informative locations (PM or MM probes only) with the same code (corresponding to array r, say), then a spatial bias is present and we adjust the value in cell (i, j, r). Let Δ be the subset of the N informative locations within the window. For each location in Δ we calculate the d values for array r, and let $\bar{d}$ be their average. The adjusted value L^a_{ijr} is given by

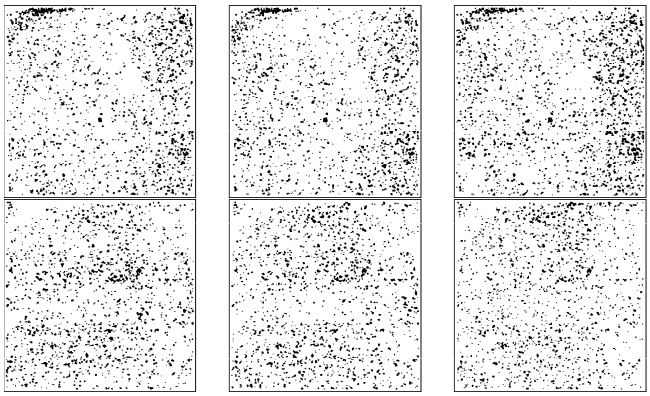
$$L^a_{ijr} = L_{ijr} - \beta_{ij}\bar{d}. \tag{5}$$

3.3 Results

Using two bias-free arrays (the first and second replicates) obtained in Section 3.1 together with only one contaminated array (any of the arrays used in Section 2) we

apply LPE, repeating the procedure separately for the three contaminated arrays. Figure 3 shows the remaining spatial biases. Although a number of locations remain that have unusually high or unusually low values these are not concentrated in regions of the arrays or have defined shapes.

Figure 3: Remaining spatial flaws in three HG-U133 Plus 2.0 arrays: (a) GSM46959 (left), (b) GSM76563 (center) and (c) GSM117700 (right). The layout is as in Figure 1.



4 Conclusions

The problem of identifying and reducing spatial biases has scarcely been adressed in the literature. We have found many examples of microarray experiments where these biases are present (regardless of the chip or design type), so the problem is not uncommon. Some methods have been proposed to identify and reduce these biases using replicate arrays, however these methods do not work in the absence of replicates. We have proposed a new method to identify spatial biases without replication, and extended an existing method in order to reduce the effects of these flaws.

References

Arteaga-Salas, J.M., Harrison, A.P., and Upton, G.J.G. (2008). Reducing spatial flaws in oligonucleotide arrays by using neighbourhood information. *Statistical Applications in Genetics and Molecular Biology*, Accepted.

Langdon, W.B., Upton, G.J.G., da Silva Camargo, R. and Harrison, A.P. (2008). A survey of spatial defects in Homo Sapiens Affymetrix GeneChips, IEEE/ACM *Transactions on Computational Biology and Bioinformatics*, Submitted.

Suárez-Fariñas, M., Pellegrino, M., Wittkowski, K.M. and Magnasco, M. (2005). Harshlight: a "corrective make-up" program for microarray chips. *BMC Bioinformatics*, **6**, Art. 294.

EMERALD microarray platform comparison based on hypothesis tests under order restrictions

Florian Klinglmüller[1] and Thomas Tüchler[2]

1. Introduction

The EMERALD workshop at CAMDA has provided a dataset [4] for analysis that presents microarray measurements from two different tissue materials (liver, kidney) and two titrations in different proportions therof (1 : 3,3 : 1). The main motivation behind titration experiments [6] is an *a-priori* known relation between measurements from different conditions. One downside, however, is that the set of true positives is not known in advance. This motivates the use of a hypothesis test that tests for a monotonic trend in the concentrations of the titration design. This seems especially promising for the task of cross-platform comparisons. From this point of view testing the implied order restrictions can also be seen as a least common denominator that one would expect to be consistent over several technical conditions.

1.1. Microarray Data.

RNA material from two tissues, liver and kidney extracted from six genetically different rats was measured on three different commercial microarray platforms: Affymetrix Rat Genome 230 2.0 array, Illumina RatRef-12 array and Agilent Whole Rat Genome array. Samples are titration mixtures prepared in four different conditions: 100% liver material, 75% liver and 25% kidney material, 25% liver and 75% kidney material and 100% kidney material. For each condition and individual animal three technical replicates are available.

1.2. Notation.

Below we have adopted the following notations and indices. We will generally refer to observed $\log_2$ expression measures by the letter y using the following indices: For the four conditions we use index $j = L, M1, M2, K$ with the labels referring to: 100% liver material (L), 75% liver and 25% kidney material (M1), 25% liver and 75% kidney material (M2), and 100% kidney material (K), respectively. Technical replicates will have index k, $(k = 1, 2, 3)$, genes $g = 1, ..., N$ where N is the total number of genes and animal $i = 1, ..., 6$.

[1]Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria, float@lefant.net

[2]WWTF Chair of Bioinformatics, BOKU University, Muthgasse 18, 1190 Vienna, Austria

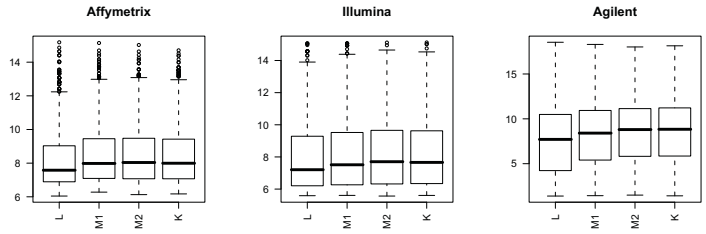


FIGURE 1. Boxplots of mean $\log_2$ expression values for the 4 tissue groups. Thick line in the box represents median, box interquartile range and whiskers 2.5 times the interquartile range, points signify values exceeding these. Consistent over all platforms pure liver samples have lower expressions in the median than pure kidney samples. This is indicative for differing total to messenger RNA concentrations.

1.3. **Exploratory analysis of unnormalized data.** In this section we present some aspects of the data uncovered by exploratory analysis of the unnormalized data. By looking at the distributions of observed expressions in the different tissue preparation (L,$M1$,$M2$,K) we can observe that the median observed signal in liver is considerably lower than in kidney. Figure 1 shows boxplots of the average $\log_2$ expression values in the 4 groups, for each platform. This observation is indicative for different total to messenger RNA proportions in the two tissues. More importantly it is relevant to the analysis of monotonicity, since using unnormalized signals we expect a tendency to upward trends.

1.4. **Tests for monotonic trend.** Monotonicity of the observed expression values in the order implied by the titration design, was assessed using the Barlow test statistic $\overline{E}_g^2$ ([1] and [8]). [5] have recently discussed this statistic in the context of microarray experiments and developed appropriate multiple testing procedures. A major advantage of this approach is that it is based on isotonic regression and does not depend on assumptions about the functional form of the trend (e.g. linear). For each gene g and animal i we have 3 observations for each of the four conditions (L,$M1$,$M2$,K) which induce a natural order. We test the null hypothesis of equal mean expression levels $\mu_{j,g}$

$$H_{0,g} : \mu_{L,g} = \mu_{M1,g} = \mu_{M2,g} = \mu_{K,g}, \tag{1}$$

against the ordered alternatives

$$H_{1,g}^{up} : \quad \mu_{L,g} \leq \mu_{M1,g} \leq \mu_{M2,g} \leq \mu_{K,g}, \tag{2}$$

$$H_{1,g}^{down} : \quad \mu_{L,g} \geq \mu_{M1,g} \geq \mu_{M2,g} \geq \mu_{K,g}, \tag{3}$$

with at least one strict inequality. In order to adress the variance structure implied by the experimental design (3 technical replicates per animal and condition, 6 animals) we calculate test statistics seperately for each of the 6 animals and then combine the resulting p-values across the 6 animals. The $\overline{E}_{ig}^{2up}$ statistic for gene g and animal i is the ratio of the sum of squares explained by isotonic regression means, assuming an upward trend, against the sum of squares explained by the equal means assumption. The null hypothesis is rejected for large values of $\overline{E}_{ig}^{2up}$. Similarly, the $\overline{E}_{ig}^{2down}$ is calculated accordingly using isotonic regression assuming a downward trend. Thus, for each gene and animal we obtain two one sided p-values p_{ig}^{up} and p_{ig}^{down} using the permutation null distribution. To obtain overall one sided test statistics over all 6 animals, we combine the one sided p-values using the inverse normal combination function

$$p_g^{C,up} = 1 - \Phi\left(\frac{1}{\sqrt{N}} \sum_i \Phi^{-1}(1 - p_{ig}^{up})\right), \tag{4}$$

where Φ is the standard normal distribution function and Φ^{-1} its corresponding quantile function. $p_g^{C,down}$ is defined by analogy. Under the null hypothesis of equal expression levels across the four conditions and 6 animals $p_g^{C,up}$ and $p_g^{C,down}$ are uniformly distributed and thus conservative one-sided p-values for the corresponding null hypotheses. A two-sided p-value is given by $p_g^C = 2\min(p_g^{C,up}, p_g^{C,down})$. The directional decision is then made by choosing the smaller of the two p-values. To adjust for multiplicity we use the Benjamini-Hochberg [2] procedure controlling the false discovery rate (FDR), which has also been shown to control the mixed-directional FDR [3]. Computation of test statistics and p-values are based on the library `IsoGene` provided by [5] for the `GNU R` statistical programming language [7].

1.5. **Results.** In the following analysis we used data normalized by either quantile or median baseline normalization to reflect the typical practice of statistical analysis. Simple order assesment of the observed expression means $\overline{y}_{jg}$ in the different conditions (j=L, $M1$, $M2$, K) pooled over technical replications and animals as done in [9] suffers the problem that e.g. assuming 4 *i.i.d.* random variables each of the $4! = 24$ possible order combination will be equally likely, with two of them corresponding to monotonicity, hence resulting in a 1/12 chance for a strictly monotone trend. Additionally, by looking only at genes for which the absolute mean difference $|\overline{y}_{Lg} - \overline{y}_{Kg}|$ is above a certain threshold the probability to observe monotonicity by chance increases even more and reaches 50% as the threshhold for the difference increases. To improve on these shortcomings we use hypothesis tests outlined above that control a multiplicity adjusted Type I error rate.

1.6. **Across platform consistency.** Since approximation of empirical null distributions using sample label permutations is computationally intensive

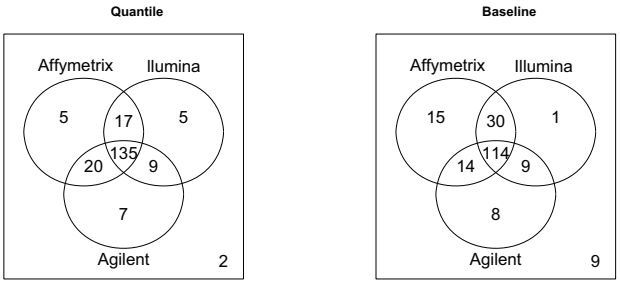


FIGURE 2. Venn diagramms showing the overlap of significant rejections between the three platforms. Left panel shows results from quantile normalized data, right panel from baseline normalized data.

[3] for the time being, we only report results for a randomly selected subset of 200 genes (out of the approx. 6900 genes common to all three platforms) using 1000 permutation iterations. Across platform consistency was assessed based on the two-sided p-values p_g^C. Figure 2 shows a venn diagram summarizing the overlap of rejections in the three different technologies with the left panel referring to quantile normalized data and the right to baseline normalized data. The proportion of genes significant in all three technologies is around 70%.

1.7. **Difference between normalization methods.** The different locations of the observed expression distributions (see Figure 1) suggests that without normalization we should expect a tendency for upward trends. Both normalization methods are able to remove any observable differences in location and scale between the 4 groups (data not shown). The results from hypothesis testing differ between the two methods. For example the following table shows a cross tabulation of results from the Agilent platform:

	Quantile	
Baseline	Sign.	Non. Sign
Sign.	137	8
Non. Sign.	34	21

For six hypotheses we observed that the monotonic trend is significant for both normalization methods but the direction changed.

[3] We estimate more than 200 hours of computation time for all of the genes using 10000 permutations with the hardware we have available

2. Discussion

The applied hypothesis tests detected a monotonic trend in the expression levels in a very large proportion of investigated genes. However, normalization methods typically rely on the assumption that the majority of genes are not differentially expressed which may be one reason for our finding that for a considerable proportion of genes the test results do not coincide for the two normalization procedures. In a next step we will investigate the consistency accross platforms as well as across normalizations methods for all genes that are present in all three platforms.

References

[1] Richard E. Barlow. *Statistical Inference Under Order Restrictions*. John Wiley and Sons Ltd, 1972.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289-300, 1995.

[3] Yoav Benjamini and Daniel Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100:71–81, 2005.

[4] W. Ligget, R. Peterson, and M. Salit. Technical vis-à-vis biological variation in gene expression measurements. 2008.

[5] D. Lin, Z. Shkedy, D. Yekutieli, T Burzykowski, H. Gaehlmann, A. Bondt, T. Perera, T. Geerts, and L. Bijnens. Testing for trends in dose-response microarray experiments: a comparison of several testing procedures, multiplicity and resampling-based inference. *Statistical Applications in Genetics and Molecular Biology*, 6:Article26, 2007.

[6] MAQC consortium. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–1161, 2006.

[7] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[8] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. John Wiley & Sons Inc, December 1988.

[9] R Shippy, S. Fulmer-Smentek, R. Jensen, W. Jones, P. Wolber, C. Johnson, Pine P., C. Boysen, Xu Guo, E. Chudin, Y.A. Sun, J Willey, J. Thierry-Mieg, D. Thierry-Mieg, R. Setterquist, M Wilson, A. Bergstrom Lucas, N. Novoradovskaya, A. Papallo, Y. Turpaz, S. Baker, J. Warrington, L. Shi, and D. Herman. Using rna sample titrations to assess microarray platform performance and normalization techniques. *Nat Biotech*, 24:1123–1131, 2006.

Exploiting the EMERALD mixture design for model based microarray platform comparisons by Bayesian inference of technical and biological variance components

Thomas Tüchler* Florian Klinglmüller† Peter Sykacek* David P Kreil*

October 23, 2008

Abstract

With the difficulty of constructing a biologically relevant 'gold standard', an evaluation of the performance of different microarray platforms remains a challenge. The EMERALD contest data set examines known tissue mixtures and a biological variance component *vis-a-vis* technical variations, allowing a comparative analysis of the different gene expression profiling platforms studied. We here introduce and apply a fully Bayesian model for the inference of the variance components which explicitly exploits the tissue mixtures featuring in the EMERALD experiments. The model permits an assessment of each platform's ability to detect biological variation. We observed intensity dependent differences specific to each platform and determined that biological variance amounts to about 30% of the signal variance in this data set.

Introduction

Evaluating microarray performance across platforms is not straightforward. Challenges already arise, at the point where one has to decide, which measurements are supposed to correspond to each other. Although targeting the same organism, the three platforms in the contest data set differ considerably in the number of measurements they provide (from 22,500 in Illumina to 41,000 in Agilent). With genome annotations constantly evolving, it is thus necessary to limit an evaluation to a set of genes, that are measured by all platforms, and for which the underlying mRNA sequence is reasonably well established. With probe design being crucial for microarray performance [1], it needs to be assured that the reporters compared to each other were designed for the same, existing and well defined mRNA species.

Focussing on the evaluation task itself, current approaches range from determination of technical precisions to elaborate spike-in experiments. While the former approach does not require any external reference, its ability to determine accuracy is limited. On the other hand, establishment of spike-in experiments allowing extrapolation to complex mixtures in real world samples is still a field of ongoing research [2–5].

Pursuing an approach with known mixtures of complex samples, as performed in the EMERALD experiments to be analysed in this contest, offers an interesting alternative. RNA samples from a widely used model organism, *Rattus norvegicus*, provide the realistic setting for this study and the mixtures enable model based analysis of the data [6]. For *quantitative* evaluation, we resort to the fact that the biological replicates, *i. e.,* the different rats in this study, contribute variation to gene expression levels; be that due to mutations, epigenetic or environmental effects in their live history. Being able to reliably detect these biological differences, despite the technical noise component blurring the

*WWTF Chair of Bioinformatics, BOKU University Vienna, Muthgasse 18, 1190 Vienna, Austria. The authors can be contacted via email at firstname.lastname@boku.ac.at, *e. g.,* at thomas.tuechler@boku.ac.at.

†Medical Statistics and Informatics, Spitalgasse 23, Medical University of Vienna, 1090 Vienna, Austria, float@lefant.net

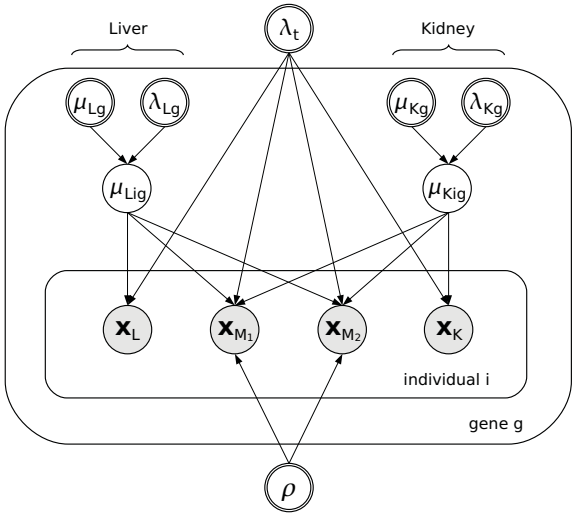


Figure 1: Directed acyclic graph of the variance component model exploiting tissue mixture ratios.

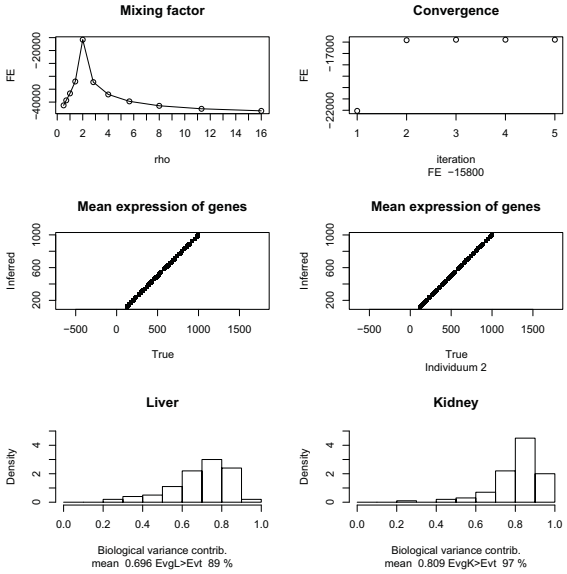


Figure 2: Simulation study validating Variational Bayes inference of the model: recovery of a known mixture ratio $\rho = 2$, convergence of the lower bound on the model evidence after 5 iterations already, recovery of overall and animal gene means, and recovery of known biological variance contributions of 67% and 80%, respectively.

measurements, can therefore serve as a criterion for a successful experiment. We hence devised a Bayesian model to infer the amount of variance that can be explained biologically, as well as the uncertainty in the mRNA contents of the tissue mixtures.

In summary, we assessed the three platforms ability to extract biological differences using a model dedicated to the particular mixture design in this data set.

Methods

Preprocessing

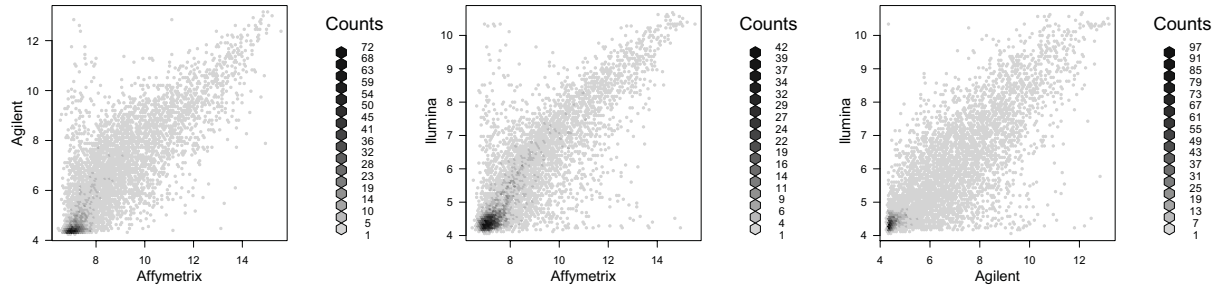
A subset of reporters designed against unique and well established mRNA species was created by filtering for 'NM' RefSeq identifiers common to all three platforms (`www.ncbi.nlm.nih.gov/RefSeq/`). Following McCall and Irizarry in their spike-based platform comparison [5], Affymetrix data were considered as summarised by RMA. However, we expanded on their preprocessings by investigating baseline, quantile and VSN [7] normalization individually for all three platforms. In addition, we also investigated reannotated [8] and GCRMA preprocessed [9] Affymetrix data in conjunction with VSN and PLM detrending [10] as described by Hüttel, Kreil *et al.* [11].

Inferring variance components

Biological and technical variance components were quantified in a fully probabilistic way. To this end, r replicated expression measurements $\mathbf{x}_{Tig}$ for tissue T, individual rat i and gene g were modelled with biological and technical precision, λ_{Tg} and λ_t; the latter one being shared by all genes [12]. Introducing a tissue mixture ratio $\rho = \frac{[\mathrm{mRNA}_K]}{[\mathrm{mRNA}_L]}$ to account for unequal mRNA concentrations in liver and kidney samples [6, 13], the hierarchical model,

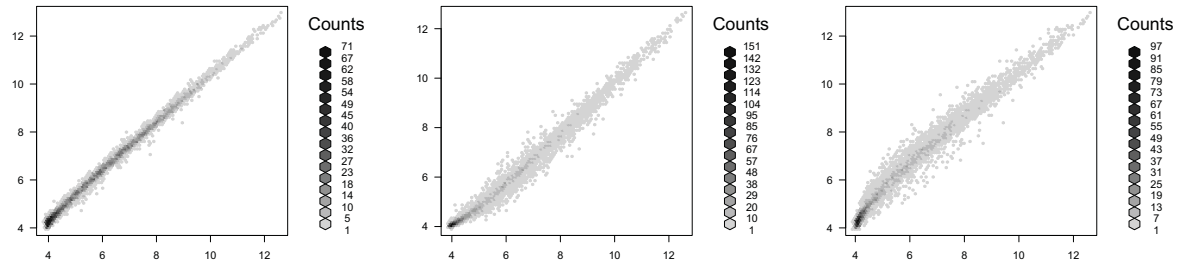
$$\begin{aligned}
\mathbf{x}_{Lig} &\sim \mathcal{N}\left(\mu_{Lig}, \lambda_t^{-1}\right) \\
\mathbf{x}_{M_1ig} &\sim \mathcal{N}\left(\rho_{M_1} \cdot \mu_{Lig} + (1-\rho_{M_1}) \cdot \mu_{Kig}, \lambda_t^{-1}\right) \\
\mathbf{x}_{M_2ig} &\sim \mathcal{N}\left((1-\rho_{M_2}) \cdot \mu_{Lig} + \rho_{M_2} \cdot \mu_{Kig}, \lambda_t^{-1}\right) \\
\mathbf{x}_{Kig} &\sim \mathcal{N}\left(\mu_{Kig}, \lambda_t^{-1}\right) \\
\mu_{Lig} &\sim \mathcal{N}\left(\mu_{Lg}, \lambda_{Lg}^{-1}\right) \\
\mu_{Kig} &\sim \mathcal{N}\left(\mu_{Kg}, \lambda_{Kg}^{-1}\right) \ ,
\end{aligned}$$

with uninformative Gaussian priors on the gene means μ_{Lg} and μ_{Kg}, and uninformative Gamma



(a) Affymetrix *vs.* Agilent (b) Affymetrix *vs.* Illumina (c) Agilent *vs.* Illumina

Figure 3: Cross platform concordance. Signal responses differ substantially between platforms, especially in the lower intensity regions. The plots show baseline normalised gene expressions in liver samples, averaged over animals and replicate arrays for each of the three platforms on subset of 6111 high confidence comparable genes.



(a) 5B-2 *vs.* 5B-1 (b) 5B-2 *vs.* 5B-3 (c) 5B-1 *vs.* 5B-3

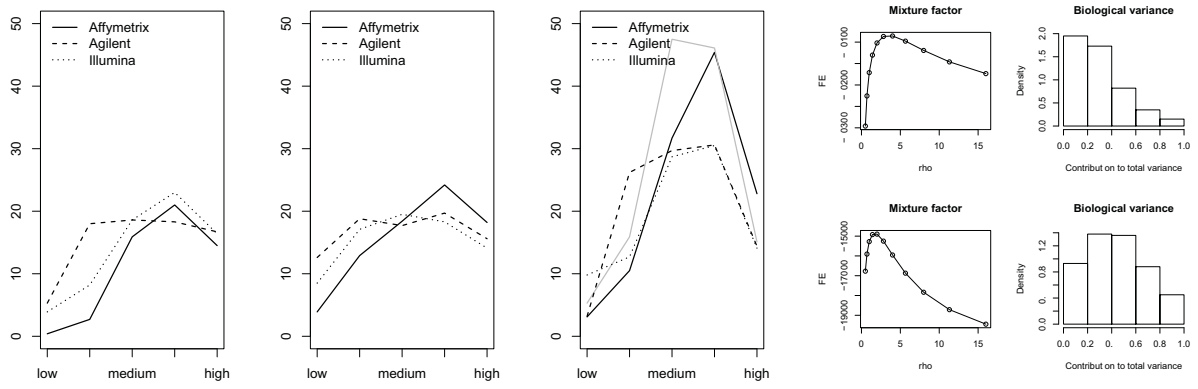
Figure 4: Outlier slides. While the first two technical replicates, here 5B-2 and 5B-1, align well in this example from the Agilent dataset, the third one, 5B-3, behaves differently with outliers affecting all platforms.

priors on the biological precisions λ_{Lg} and λ_{Kg}, was then inferred applying Variational Bayes [14]. Besides fully exploiting the mixture design and providing full posterior distributions over the model parameters, this approach is computationally feasible even for microarray sized data sets. Also note, that sharing a technical variance distribution for all genes, renders a subsequent regularization of the technical variance redundant [15] and makes estimates more robust [16]. A directed acyclic graph representation of the model and simulation studies validating its performance, are depicted in Figs. 1 and 2.

Results

Inspection of the exploratory plots in Fig. 3 reveals considerable between-platform variation ($R^2 \sim 0.65$). This is especially the case for the low intensity measurements. Affymetrix and Illumina spread low signals across a wider intensity range as compared to Agilent. Within-platforms, obvious outlier slides, as depicted in Fig. 4, contributed substantial variation. For Affymetrix, array 3B-3 was recognized as such an outlier. For the Illumina data, outliers comprised experiments with low cRNA yield (3 of 6), but even more so plate location 3 (4 of 6) and hybridization data 10/06/08 (5 of 6). For Agilent the third replicate series (hybridization names '. . . -3') conducted by operator 'A' matched the others poorly. Interestingly, this operator also corresponded to an increased 'AmpLabelingInputMass'. Since removal of outlier arrays improved the evaluation statistics, we here only report the results from these cleaned data sets.

Evaluation of how often the biological variance exceeds the technical one in Figs. 5(a) to (c) show that in the lower intensity ranges Agilent reports the most variation, with 5-10% of genes having greater



(a) Summary baseline (b) Summary quantile (c) Summary VSN (d) Inferred distributions

Figure 5: Biological variance. Panels (a) to (c) plot, different normalization methods, the percentage of genes that have a higher biological than technical variance component, as a function of signal intensity, providing a direct comparison of all three platforms. The grey line in panel (c) indicates Affymetrix data from the alternative GCRMA-VSN-PLM preprocessing. Panel (d) depicts how the inferred mixture factor ρ and the distribution of biological variance contributions differs between genes of low (top row) and high intensity (bottom row); shown for Illumina liver measurements.

biological than technical variance. Affymetrix, on the other hand, has almost no biological variance detectable in baseline corrected low intensity data. More elaborate normalization methods, like quantile or VSN normalization, however, increase this feature to 5% for Affymetrix and to more than 10% for Agilent and Illumina. Intriguingly, the situation changes for medium and and high intensity data. While all three platforms identify comparable biological differences between individuals in the medium intensity ranges – 20% of genes with at least half of the total variation explained biologically – Affymetrix finds up to 45% of such genes in the higher intensity segment of VSN normalized data. Additional preprocessing efforts (GCRMA, PLM) lead to similar results in medium range Affymetrix data already. Eventually, regardless of platform and normalization, the extractable biological variances drop again in the highest intensity segment, which could indicate saturating signal responses. On average, about 20% of the genes were identified with a biological variance component above the technical noise level, explaining about 30% of the overall variance.

The posterior distribution of the mixture parameter ρ provides a further indicator for how well the data capture the titration design[1]. The less the measurements reproduce a monotonous trend from pure liver to pure kidney samples, the more uncertainty about ρ is obtained. If liver and kidney specific genes can not be distinguished, the ratio defining the mixtures M_1 and M_2 becomes irrelevant. Panels on the left in Fig. 5d exemplify that for low and high intensity Illumina data. Like Liggett et al. we observed more mRNA in the kidney total RNA samples ($\rho > 1$). We also want to emphasize, that the better the mixture design is reflected by the data, the more information about individual gene expression values can be gained from the mixture samples. This will provide more power for inferring the biological variance component. Our evaluation criterion is thus directly linked to the preservation of the mixture design.

Apart from these technical observations, we also found that the liver samples show more variation than those from kidney. With no obvious biological explanation and variations in gene expression having functional implications [17], this remains a potential topic for further biological investigations.

[1]Note, that our model considers both the known ratios of total RNA mixtures and the *unknown* mRNA concentrations within the liver and kidney total RNA samples (*cf.* Shippy et al. for detailed derivations).

Discussion

The ability to determine biological differences despite inherent technical fluctuations, is a key requirement for any microarray experiment. Assessing the success of an experiment by the number of genes for which the presumed biological signal exceeds the technical noise thus seems a sensible idea. We have exploited the biological replicates in this study to quantify how much natural variation between individual animals can be detected using different microarray platforms probing the *same* genes. We put a strong emphasis on comparing only a subset of reporters, that are not likely to suffer from changes in genome annotation, or uneven fractions of low confidence gene predictions within the three microarray platforms. Analysis was therefore based exclusively on reporters targeting well curated, unique reference sequences.

To establish biological variance as a measure for platform performance, we have stipulated that a) there are in fact differences between animals and b) that these animal effects are not confounded with other experimental or normalization effects; *i. e.,* that a randomised study design enables valid inference of the biological variance component. Based on that, we devised a Variational Bayes model accommodating the particular mixture design of the contest data set. At this point we wish to highlight, that inference within the Variational Bayes framework, though not yet widely used in the field, provides a range of nifty features: Full posterior distributions for all variables are obtained in a single inference step, which is particularly useful for modelling mixtures in titration experiments [6, 13]. Moreover, the Variational Bayes algorithm is computationally efficient and scales well even for microarray data. Eventually, a lower bound on the model evidence can be derived by the means of Variational Bayes, providing a direct measure for convergence of the algorithm and supporting straightforward model comparisons.

In contrast to previous studies of biological *versus* technical variance [18], we find that that the biological component in this experiment is on average about a third of the technical noise. Clearly, biological variation depends on differences in genetic background and individual life history. The smaller the genetic differences, the better the experimental environment is controlled, the more subtle the biological variation will be; in turn increasing requirements for a successful microarray experiment.

Conclusion

Variation between individual rats in the EMERALD data set is small in relation to the technical measurement noise. Typically about 20% of the genes show biological variance exceeding technical fluctuations. Intriguingly, the three platforms' ability to detect this biological component differs remarkably with preprocessing and signal intensity.

Acknowledgements

The Boku Bioinformatics group acknowledges support by the Vienna Science and Technology Fund (WWTF), the Austrian Centre of Biopharmaceutical Technology (ACBT), Austrian Research Centres Seibersdorf (ARCS), and Baxter AG. Thomas Tüchler acknowledges partial support by the Austrian Gen-AU program.

References

[1] G. G. LeParc, T. Tüchler, G. Striedner, K. Bayer, P. Sykacek, I. Hofhacker, and D. P. Kreil. Model based probe set optimization for high-performance microarrays. *Nucl. Acids Res., in revision*, 2008.

[2] SE. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16, 2005.

[3] A. R. Dabney and J. D. Storey. A reanalysis of a published affymetrix genechip control dataset. *Genome Biology*, 7(3):401, 2006.

[4] R. A. Irizarry, L. M. Cope, and Z. Wu. Feature-level exploration of a published affymetrix genechip control dataset. *Genome Biology*, 7(8): 404, 2006.

[5] M. N. McCall and R. A. Irizarry. Consolidated strategy for the analysis of microarray spike-in data. *Nucleic Acids Res*, 2008.

[6] R. Shippy, S. Fulmer-Smentek, R. V. Jensen, W. D. Jones, P. K. Wolber, C. D. Johnson, P. S.

Pine, C. Boysen, X. Guo, E. Chudin, Y. A. Sun, J. C. Willey, J. Thierry-Mieg, D. Thierry-Mieg, R. A. Setterquist, M. Wilson, A. B. Lucas, N. Novoradovskaya, A. Papallo, Y. Turpaz, S. C. Baker, J. A. Warrington, L. Shi, and D. Herman. Using rna sample titrations to assess microarray platform performance and normalization techniques. *Nat Biotechnol*, 24(9):1123–31, 2006.

[7] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

[8] M. Dai, P. Wang, A. D. Boyd, G. Kostov, and B. Athey. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucl Acids Res*, 33(e175), 2005.

[9] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez Murillo, and F. Spencer. A model based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc*, 99(909), 2004.

[10] B. Bolstad. *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.* PhD thesis, University of California, Berkeley, 2004.

[11] Bruno Huettel, David P. Kreil, Marjori Matzke, and Antonius J. M. Matzke. Effects of aneuploidy on genome structure, expression, and interphase organization in arabidopsis thaliana. *PLoS Genet*, 4(10):e1000226, Oct 2008. doi: 10.1371/journal.pgen.1000226.

[12] L. Wernisch, S. L. Kendall, S. Soneji, A. Wietzorrek, T. Parish, J. Hinds, P. D. Butcher, and N. G. Stoker. Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics*, 19(1):53–61, 2003.

[13] W. Liggett, R. Peterson, and M. Salit. Technical vis-à-vis biological variation in gene expression measurements. *preprint*, 2008.

[14] T. Leen, editor. *A Variational Bayesian Framework for Graphical Models*, Cambridge, MA, 2000. NIPS 12, MIT Press.

[15] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

[16] D. P. Kreil and D. Wild, editors. *Estimating Variance Components Using Variational Bayes*, 2008. Probabilistic Modelling in Computational Biology.

[17] L. Lopez-Maury, S. Marguerat, and J. Bahler. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat Rev Genet*, 9(8):583–93, 2008.

[18] D. Stivers, J. Wang, G. Rosner, and K. Coombes, editors. *Organ-Specific Differences in Gene Expression and UniGene Annotations Describing Source Material.*, 2002. Critical Assessment of Microarray Data.

A Data Transformation Ontology for Microarrays

James Malone, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

The diversity in microarray experiment designs and applications requires that a large number of pre-processing approaches are available. In order to facilitate unambiguous and consistent descriptions of experimental data transformation the development of a 'normalisation and transformation ontology' (NTO) has been undertaken, which we describe in our talk. This provides a means to conceptualize and classify the approaches used, describe relationships between these concepts and store these in a machine readable form. Such a representation can offer a useful checking mechanism to ensure that data is correctly modelled as well as a more powerful querying mechanism. The NTO has been developed as part of the Ontology for Biomedical Investigations (OBI), a large, multi-national, collaborative community development project. A Beta version of the ontology is now available from http://obi-ontology.org/page/Main_Page.

Keynote

Multiple Testing on the Graph of Gene Ontology

Jelle Goeman, Leiden Iniversity Medical Center, The Netherlands

Gene set testing methods often test gene sets derived from Gene Ontology. When testing the whole graph of Gene Ontology (GO) it is important to correct for multiplicity, and the question arises naturally how we can make use of the graph structure of GO for multilicity correction. We propose a multiple testing method, called the focus level procedure, that preserves the graph structure of Gene Ontology (GO). The procedure is constructed as a combination of a Closed Testing procedure with Holm's method. It requires a user to choose a “focus level” in the GO graph, which reflects the level of specificity of terms in which the user is most interested. This choice also determines the level in the GO graph at which the procedure has most power. The procedure strongly controls the family-wise error rate without any additional assumptions on the joint distribution of the test statistics used. We also present an algorithm to calculate multiplicity-adjusted p-values. Because the focus level procedure preserves the structure of the GO graph, it does not generally preserve the ordering of the raw p-values in the adjusted p-values.

Effect of Single Nucleotide Polymorphism (SNP) in Affymetrix probes

Olivia Sanchez-Graillet osanch@essex.ac.uk
William B. Langdon wlangdon@essex.ac.uk
Andrew P. Harrison* harry@essex.ac.uk

Department of Mathematical Sciences and Department of Biological Sciences,
University of Essex, Wivenhoe Park, Colchester, UK, CO4 3SQ
* Corresponding author

Abstract
We have performed a study on the impact of Single Nucleotide Polymorphisms upon Affymetrix probes on 3' GeneChips. We wished to explore whether the existence of a SNP always causes a probe to behave differently to other probes within a probeset. We have focussed on those probes which map uniquely to a single exon, and have little overlap with other transcripts. We use large surveys of GeneChips obtained from the Gene Expression Omnibus to derive correlations between groups of probes which map to the same exon. The resulting group of correlations are used to identify probes which appear as outliers with respect to the other probes mapping to the exon. We analysed whether the outliers result from the existence of SNPs.

1 Introduction

Affymetrix GeneChip technology provides multiple measures of the expression level for each gene. Each probe is a 25-nt oligomer (25mer) and each probeset, designed to represent a different gene transcript, typically consists of eleven perfect match (PM) probes as well as corresponding mismatch (MM) probes. It is widely assumed that multiple probes from within the same probeset measure the same thing but there are a number of probesets which contain probes which behave inconsistently with the rest of the probeset. We have begun a study to unravel the mechanisms which act to cause some Affymetrix Probes to behave as outliers with respect to other probes. Some probes are either unresponsive (no hybridization signal) or invariant (same hybridization signal) across many observations. We have also discovered that probes containing a contiguous run of 4 or more guanines are particularly prone to being outliers – we associate this effect with the formation of G-quadruplexes occurring on the surface of a GeneChip. Further, we find that probes on the edge of a GeneChip are correlated with each other, rather than detecting the biological signal for which they were chosen. In this study we have examined the impact of SNPs, and whether they are responsible for outliers.

The study of the effect of SNPs on GeneChips has been of great interest in recent years. Kumari et al. (2007) analysed the intensity distribution of probes considering PM-MM difference. They suggest that different properties are measured by the two types of probes and recommend that any analysis of GeneChips should consider the location of SNPs in order not to miss valuable information. Alberts et al. (2007) demonstrated that genetic variation affects hybridization of probes and that this might misguide the interpretation of data from individual genes, even if only a single probe is affected. However, they determined that not every SNP causes a difference in hybridization. Hughes et al. (2001) also argue that when a SNP is located at any of the extremes of a probe it may have little or no effect on hybridization.

2 Materials and methods

We are developing a pipeline which analyses tens of thousands of Affymetrix GeneChips (Sanchez-Graillet et al., 2008). Our pipeline brings together unique mapping of probes, quality control analysis on each GeneChip and data-mining signal intensities across many experiments. For the present study, we analyse the data contained in ten designs of Human arrays. We identify probes that uniquely map to exons (Sanchez-Graillet et al. 2008) and which also contain SNPs. Having this information, we use the heatmaps of exons obtained by our pipeline to select those SNP-probes which are outliers.

We have chosen 3' GeneChips, rather than exon arrays, to study exons for two main reasons: i) GEO contains many thousands of 3' GeneChips, but only a few hundred exon

arrays, and so we can derive more robust correlation values; ii) on an exon array there are only four probes per exon, whereas on a 3' GeneChip there are some exons for which there are several tens of probes mapping to the same exon – this increase naturally facilitates the detection of outliers.

We briefly describe our pipeline to get correlation heatmaps and then our method to locate SNPs.

2.1 Pipeline for obtaining heatmaps

We have downloaded tens of thousands of Affymetrix GeneChips from the Gene Expression Omnibus (Barrett et al. 2007). We are able to identify spatial flaws in individual GeneChips (Upton et al 2005, Arteaga-Sala et al.,2008, Langdon,2008) and this leads us to blank out signals from a fraction of each chip. We group all probes taken from the same exon together, and we calculate the correlations between each of the probe-pairs - we transform their intensities onto a log2 scale and correlate the signals across all examples of a given chip design. All the pair-wise probe correlations for each exon are collated, including the correlations between PM probes, MM probes and PM-MM probes, into a matrix which is colour-code according to the correlation value. The original correlations values are rounded and multiplied by 10, so that we express the correlations as integers. Heatmaps are symmetrical matrices in which the diagonal represents the perfect correlation of each probe with itself (correlation with value 10).

Heatmaps were created for every Ensembl exon (release 48) with unique Affymetrix probes aligning in the sense direction of the exon. The groups of probes that align to the same exon are expected to show concordance in their expression. Figure 1 shows a heatmap of an exon. The numbers on the left are the correlation values and the numbers on the right are the standard deviations. The numbers at the bottom indicate the number of bases between two consecutive probes.

2.2 SNP identification

The information about SNPs was downloaded from the Ensembl (Hubbard et al., 2007) database with Biomart (Durinck et al., 2005). This information includes the location of SNPs on the transcript. For our analysis, we took the SNPs which are located only in 3' UTR, 5' UTR, and coding regions. We considered all types of variation (i.e. , SNPs, indels, mixed, etc.). Our method to locate SNPs consists of the following steps:

- Identify the Ensembl exons which contain SNPs according to the transcript information and chromosomic positions of the SNPs;
- Select only exons with probes that uniquely map to them (here called "unique exons");
- Only unique exons with more than four probes are selected. The positions of the SNPs on the probes mapping uniquely to each exon are obtained.

To select unique exons with SNP-probes which are outliers it was necessary to:

- Obtain the correlation data corresponding to each unique exon;
- Compare the median of the overall correlation matrix with the meadians of each SNP-probe;
- If the difference between the two medians >= 0.15 then the probe is considered to be an outlier.

We chose to compare medians as a way to determine outliers because they represent how far a set of values corresponding to the probe being analysed is from the rest of the values of the rest of probes mapping to a particular exon. After trying different threshold values, 1.5 was found to be the optimal threshold to determine outlier probes. The diagonal values are not included in the calculation, as these values result from comparing a probe with itself. The values of the probe to be compared are also not included. Since heatmaps can contain probes whose sequences overlap with other probe sequences (i.e. that are not independent of each other), we could not make use of non-parametric statistical tests that require independent samples (e.g. Mann-Whitney U-test).

3 Results

Overall, we looked at 59,666 SNPs distributed in unique exons of ten array designs. We found

6,782 SNPs (11.37%) in unique exons in which all probes that contain the same SNP are outliers, 2,888 SNPs (4.84%) in which not all the probes containing the same SNP are outliers, and 49,996 SNPs (83.79%) in which all probes are not outliers.

The most frequent variation found was SNP, followed by indels and mixed variation. The most common allele was C/T. If more than one SNP-probe maps to the same exon, the probes may have partial or total overlapped sequences. The SNP-probes can be from the same probeset or from different probesets. In several cases in which the probes of the same probeset are highly correlated in the PM heatmap and there is a probe that contains a SNP on position 13, this probe constitutes an outlier in the PM-MM heatmap.

Figure 1 shows an example of a heatmap of an exon which contains SNPs. Table 1 gives information about the SNPs contained in the exon.

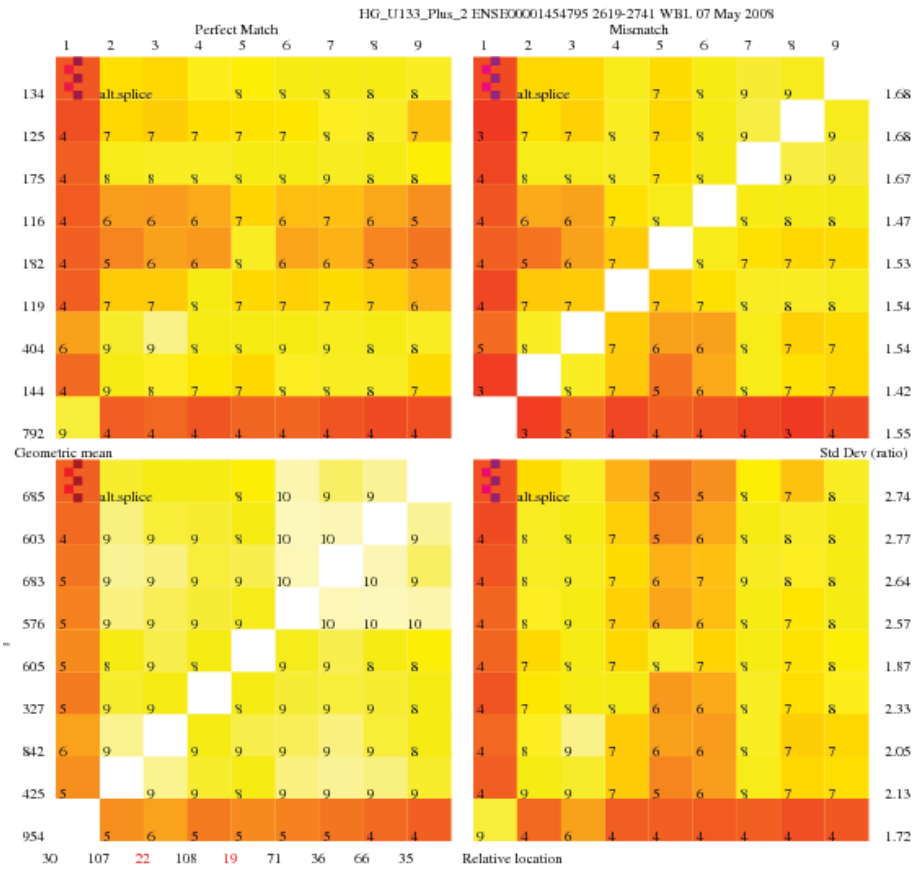


Figure 1. Exon **ENSE00001454795** contains five different SNPs. Only SNP **rs1049814**, which is in the coding region, is an outlier probe (See Table 1). The rest of the SNP-probes are in the 3'UTR and do not constitute outliers. This exon is in strand -1. Therefore, the alleles in the probe sequence are the complement of the given allele. The SNPs are in different positions on the respective probes. The MM correlations do not significantly differ from the PM correlations.

Probe position on heatmap	Probe Id	SNP position on probe	SNP Id	Biotype	Allele
1	208690_s_at-112-159	20	rs1049814	coding	A/G
4	208690_s_at-607-591	14	rs1049961	3utr	A/G
6	208690_s_at-584-673	9	rs10048	3utr	G/A
7	208690_s_at-874-1129	10	rs1050003	3utr	T/G
8	208690_s_at-327-349	15	rs7989	3utr	C/A

Table 1. Probes uniquely mapping to exon **ENSE00001454795**, which contain SNPs.

Particular SNPs can occur in more than one array, embedded in various probe constellations. For instance, the same SNP can be contained in two outlier probes in one array while in a different array, the same SNP can be contained in one outlier probe and in one no-outlier probe. Therefore, for the following analyses, we selected only one specific array –HG_U133_Plus_2- This array contains a total of 13,822 SNPs in unique exons.

A 2x2 cross-tabulation of SNPs (yes/no) and outlier probes (yes/no) resulted in a clearly not statistically significant *phi* value of –.002. Substantially, this result suggests no systematic association between the occurrence of SNPs and outlier probes in our heatmaps.

We also plotted the values of the main alleles (A,C,T,G) against the median differences (which determine whether a probe is an outlier or not) and the positions of the SNPs on the probes (from 1 to 25) against the difference of medians. Figures 2 and 3 show these plots.

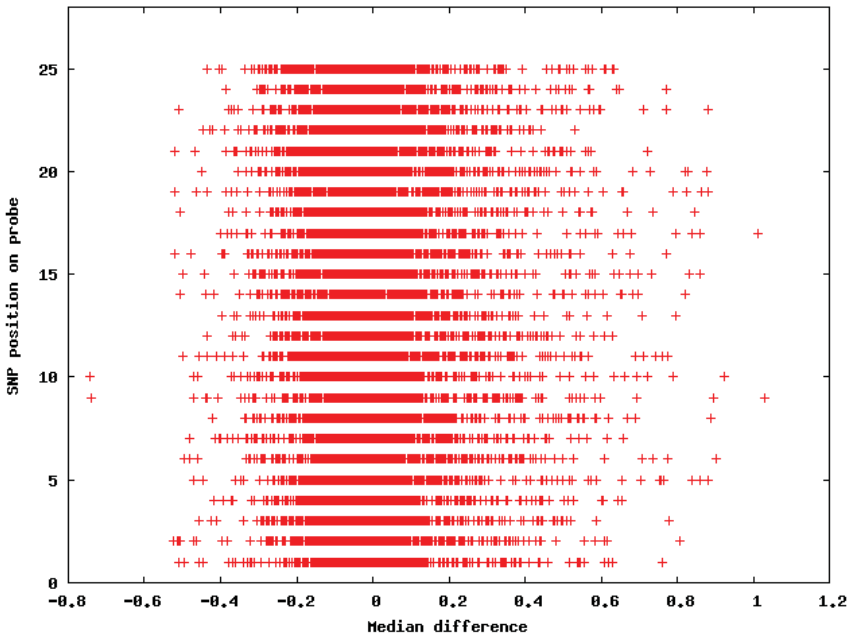


Figure 2: Scatter of median differences and positions of SNPs on probes in HG_U133_Plus_2

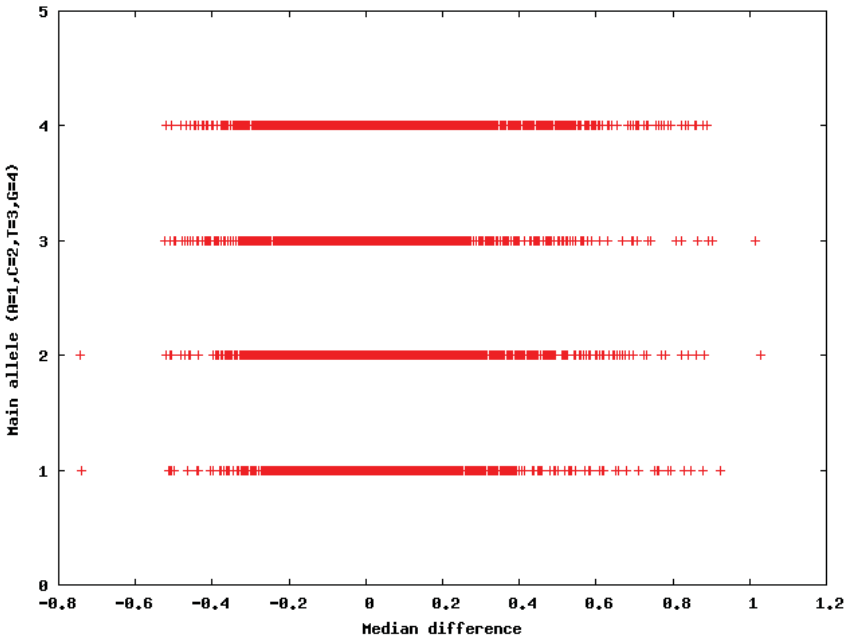


Figure 3: Scatter of median differences and main allele found in SNPs in HG_U133_Plus_2

The distributions of the alleles and positions of the SNPs (as shown in Figures 2 and 3) suggest that both, positions and alleles, do not seem to determine whether a SNP-probe also constitutes an outlier.

4 Conclusions

We have observed the effect of SNPs in correlation heatmaps of exons. We have not found a common behaviour when SNPs are present in a probe. Importantly, we suggest that SNPs do not always result in outliers in groups of probes representing individual exons. Of course, our findings are relative to our specific definition of outlier probes and the way in which the location of SNPs in the genome is determined by Ensembl. However, SNPs may influence other biological events like alternative polyadenylation. The chromosomic region where SNPs are found, the position of the SNP in a probe, and the number of SNPs in a probe does not make a probe an outlier in the correlation heatmap. In future work, we will annotate exons with post-transcriptional events and SNPs. The presence of SNPs would help in the identification of diseases or other kind of gene associations.

Acknowledgments

OSG and WBL are supported by a grant from the BBSRC (BB/E001742/1).

References

Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J.P. and Jansen, R.C. (2007). Sequence polymorphisms cause many false cis eQTLs. In *PLoS ONE*, Vol. 2.

Arteaga-Salas, J.M., Zuzan, H., Langdon, W.B., Upton, G.J.G and Harrison, A. (2008). An overview of image processing methods for Affymetrix GeneChips. In *Briefings in Bioinformatics*, 9(1), 25.

Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles – database and tools update. In *Nucleic Acids Research*, 35 (DataBase issue): D760-D765.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. In *Bioinformatics*;21(16):3439-40.

Hubbard, T. J. P., et al. (2007). Ensembl 2007. In *Nucleic Acids* Vol. 35, Database issue:D610-D617.

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., et al. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. In *Nature Biotechnology*,19:342–347.

Kumari, S., Verma, L.K. and Weller, J.W. (2007). AffyMAPSDetector: a software tool to characterize Affymetrix GeneChip expression arrays with respect to SNPs. In *BMC Bioinformatics*, 8:276.

Langdon, W.B., Upton, G.J.G, Camargo, R. and Harrison, A. (2008). A survey of spatial defects in Homo Sapiens Affymetrix GeneChips. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Submitted.

Sanchez-Graillet, O., Rowsell, J., Langdon, W.B., Stalteri, M., Arteaga-Salas, J.M., Upton, G. J.G. and Harrison, A.P. (2008). Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips. In *Journal of Integrative Bioinformatics*, 5(2):98.

Upton, G.J.G. and Lloyd, C.J. (2005). Oligonucleotide arrays: information from replication and spatial structure. In *Bioinformatics*, 21(22):4162-4168.

Extending pathways with inferred regulatory interactions from microarray data and protein domain signatures

Christian Bender*, Holger Fröhlich, Marc Johannes, Tim Beißbarth

German Cancer Research Center (DKFZ), Division of Molecular Genome Analysis (B050),
Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

Email: c.bender@dkfz-heidelberg.de;

*Corresponding author

Abstract

Motivation: The goal of this study is to demonstrate a novel approach to analyze the CAMDA 2008 contest time-course microarray dataset of Affara et al. (2007) on endothelial cell apoptosis. We utilise different methotodologies such that by the combination of the different analysis approaches a picture of the underlying regulation mechanisms is created.

Background: Integration and interpretation of various data sources is of great importance to gain a complete understanding of biological systems. Databases like KEGG, InterPro and GO offer information that can be combined with network reconstruction results for high throughput genomic data. New regulatory interactions inferred from this data can guide the focus of further research and validate the role of these novel components in biological experiments.

Results: We present a combination of a network reconstruction approach with a method for functional characterisation of the involved components. In the apoptosis microarray experiment we identify the KEGG pathway *Cell Cycle* as significantly overrepresented in the set of differentially expressed genes. We extend this pathway by novel components identified by protein domain signatures and infer interactions of the components with a dynamical bayesian network (DBN) reconstruction method. A validation of the interactions is performed via a literature network constructed by Ingenuity®software.

Conclusions: By the combination of inference of a gene regulatory network and functional characterisation of the genes in the network we show how to integrate knowledge about regulatory interactions learned from experimental data into known pathway contexts.

Background

Interpretation of high throughput data from genomics or proteomics studies is a major challenge in today's research. Huge amounts of data have to be analysed statistically to reconstruct regulation or interaction patterns of biological systems from the data. After such an analysis a list of interesting genes or proteins is usually left which have to be characterised regarding their function. The KEGG database (Kanehisa et al., 2008) offers gene annotation and visualises this in-

formation in pathway maps, but only annotation of about 4000 of the estimated 20000-25000 protein-coding genes is available. The Gene Ontology Consortium (2004) offers annotation for most genes, but not all genes have a known function. Geneset Enrichment Analysis can be used to determine overrepresented functions or pathways in gene lists (e.g. Beissbarth and Speed (2004); Al-Shahrour et al. (2004)), but is limited by the availability of gene annotation. We have devised a novel method to predict pathway membership of genes based only on the protein domain annotation and validated this method in simulation studies (Hahne et al., 2008; Fröhlich et al., 2008). The InterPro database (Mulder et al., 2008) offers protein-domain annotation for about 19000 genes. The use of such tools gives a closer characterisation of the interesting components but leaves out an integration of interaction patterns into known contexts like signaling pathways. Here we introduce a method that combines the result from a KEGG pathway prediction based on InterPro domain signatures and network reconstruction via DBN from microarray data (Lebre, 2007). We analyse the predictions to identify significantly overrepresented KEGG pathways and integrate novel interactions found in the microarray data into the present maps.

Methods

Data preprocessing

Microarray time-course gene expression data from Affara et al. (2007) was used. We selected interesting genes in the timecourse expression data by first normalising the raw expression values by variance stabilisation normalisation (VSN, Huber et al. (2002)) and successively analysing differential gene expression with *limma* (Smyth, 2004). Genes with a normalised intensity in the lower quartile of the observed intensity range in all timepoints were excluded from the analysis as they have noninformative expression profiles. Each pair of time points was analysed, and genes showing an FDR (Benjamini and Hochberg, 1995) smaller than 0.001 in at least one of the comparisons were taken as differentially expressed.

Predicting pathway membership

A functional characterisation of all genes on the array was performed. We examined the protein-domain signatures found in the InterPro database (Mulder et al., 2008) for each annotated gene. This information was mapped to a binary vector indicating the contained protein domains. For the human genome in the current version of the InterPro database this results in feature vectors of size 2752 for each gene. We used these domain signatures to predict the membership of each gene to specific KEGG pathways with a hierarchical classification scheme, implemented in the R-package *gene2pathway* (Fröhlich et al., 2008).

The package explicitely takes into account the hierarchical organisation of the KEGG database. We expect to have more accurate predictions on the top level of the KEGG hierarchy (*Metabolism*, *Genetic Information Processing*, *Enviromental Information Processing*, *Cellular Processes*) than at the bottom level of individual pathways. It is worth mentioning that *gene2pathway* can predict a mapping of a gene to multiple pathways at once. Furthermore, each prediction is accompanied with a confidence score between 0 and 1, which results from running the classification model within a bagging scheme.

Analysis of overrepresentation of specific pathways

An analysis of overrepresentation of predicted pathways in the set of differentially expressed genes was performed, as described in Beissbarth (2006). We defined two groups, one containing the genes which were found to be annotated in KEGG, and a second group

of genes for which additionally the domain-signature prediction was included. Fisher's Exact Test was used to assess statistical significance and a multiple testing correction with Benjamini-Hochberg's method was performed.

Network reconstruction with dynamical bayesian networks

In time-course datasets the number of observed genes p is usually much higher than the number of timepoints n. To deal with this situation, Lebre (2007) developed a method for inferring DBNs for the $p >> n$ situation that represent full-order conditional dependencies. The method is available in an R-package *G1DBN* that was used for the calculation. In this approach, inference of the DBN happens in a two step procedure. First, a preliminary edge selection based on first order conditional dependencies between the variables is performed, giving a DAG (directed acyclic graph) $\mathcal{G}^{(1)}$. The key assumption here is, that the true DAG $\tilde{\mathcal{G}}$ is a subgraph of $\mathcal{G}^{(1)}$. So in the second step, $\tilde{\mathcal{G}}$ is inferred from $\mathcal{G}^{(1)}$ by determining the full-order conditional dependencies among the edges in $\mathcal{G}^{(1)}$. We chose the set of all differentially expressed genes being predicted or known members of a particular pathway and extracted the expression profiles from the time-course data. Then the reconstruction was performed with parameters $\alpha_1 = 0.5$ and $\alpha_2 = 0.1$ and a least squares M-estimator.

Next we downloaded the *Cell Cycle* pathway *hsa:04110* with the *KEGGSOAP* R-package. The result of the network reconstruction was merged with the *Cell Cycle* KEGG pathway and compared to a literature based interaction network, which was generated through the use of Ingenuity®Pathways Analysis .

Results and Discussion

Analysis of differentially expressed genes and enrichment analysis

As described in Affara et al. (2007) for a pool of 10 individuals of HUVEC, RNA was prepared at time points 0, 0.5, 1.5, 3, 6, 9, 12 and 24h and hybridised to UniSet Human 20K gene chips. Gene expression was measured using CodeLink expression analysis software. From the 20265 genes on the array 18310 genes were kept as informative genes (see Methods). 1002 genes were found differentially expressed. The mapping of the microarray's ProbeID to the EntrezGeneID resulted in 14015 unique genes that could be analysed by *gene2pathway*. These were fed into the KEGG-pathway membership prediction (see Methods), in which InterPro domains for 10630 genes were found. 3385 had pathway memberships defined by KEGG, with 268 being differential. For 4206 genes predictions were made using the domain signatures and 353 of them were differentially expressed.

For each of the predicted pathways Fisher's Exact Test was performed to find out, whether a particular pathway was significantly overrepresented in the sets of differentially expressed genes. This was done once for the genes that were directly annotated in KEGG and additionally for those that were predicted to be a member of the pathway by their domain signature using *gene2pathway*. The results are shown in table 1.

pathname	p_1	p_2	K	DS
Cell cycle	0,0031	0,0004	22	30
Metabolism	1	0,0316	96	364
Cell Growth and Death	0,3877	0,0447	26	43
Nucleotide Metabolism	0,3159	0,0562	22	22
Insulin signaling pathway	0,3159	0,2787	2	2
...	...	...	...	...

Table 1: P-values for pathway overrepresentation: p_1 for pathway membership defined only by KEGG; p_2 for pathway membership by KEGG and domain signature prediction; K: number of genes found as member of the pathway in KEGG annotation, DS: number of genes assigned to a pathway by KEGG and the domain signature prediction with *gene2pathway*.

A significant overrepresentation was found for the pathways *Cell Cycle*, *Metabolism*, *Cell Growth and Death* and *Nucleotide Metabolism*, when *gene2pathway* was used for assigning the pathway to a gene. As seen in table 1, the significance for the pathways is increased when the domain signature prediction is incorporated. It also makes sense to find the pathway *Cell Cycle* and its parent map *Cell Growth and Death* overrepresented, since the microarray data originated in an apoptosis study, which is part of *Cell Death and Growth* and closely related to *Cell Cycle*. This suggests, that genes from the *Cell Growth and Death* tier show the highest activity in the time-course. For further investigation and network reconstruction exactly those differentially expressed genes, that were part of the *Cell Cycle* pathway were taken. Since *Metabolism* is a branch that can hardly be distinguished by the use of domain signatures (Hahne et al., 2008), no further examination of these pathways was performed.

Reconstruction of regulatory networks using Dynamic Bayesian Networks

The R-package *G1DBN* was used to reconstruct a DBN from the selected expression profiles. The resulting interaction network was merged with the original KEGG network, shown in Figure 1. Differential genes assigned to the *Cell Cycle* pathway by their domain signature are shown in light grey, those already contained in the KEGG annotation in dark grey. The white nodes are the remaining nodes from the original KEGG pathway which were not found to be differential. Edges found by the DBN procedure were verified in a literature interaction network generated by the Ingenuity®software.

As it can be seen in the figure, edges CCNE2 → CDKN1C, CDKN1C → MCM, RBL1 → PLK1 and MCM → PLK1 were constructed by *G1DBN* and verified in the literature network. Edges CHEK → MCM, PLK1 → CCNA2 and PLK1 → BUB1 were found as direct edges in the *G1DBN* network, and as indirect edges in Ingenuity®, verifying an interaction between the components. By identifying functional related components with the pathway prediction and using the expression data, interactions known by the literature could be reconstructed. Furthermore, new interactions are found, e.g. NASP → PLK1, NASP → MCM or UACA → BUB1. NASP is a histone binding protein which is expressed in dividing cells. UACA regulates morphological alterations required for cell growth and motility and PLK1, MCM and BUB1 are well known components in the cell cycle pathway.

Conclusions

We propose an application of an integrative approach, in which pathway membership prediction and reconstruction of DBNs are combined to predict new interactions in well known pathway contexts. We demonstrate the use of the domain signature prediction for interpretation of a microarray dataset. Instead of analysing all differentially expressed genes at once, we show how to separate the genes into sets of genes belonging together as identified by their functional characterisation. Our approach shows good concordance with literature knowledge and suggests new components as well as their putative placement in the known pathways. The methods will be made available in the R-package *gene2pathway*.

Acknowledgements

This paper is dedicated to the memory of Professor Annemarie Poustka, who was the founder and head of the Division Molecular Genome Analysis at the DKFZ. She was an inspiring scientist and a wonderful person. We thank Dirk Ledwinka for IT support. This study was supported by the Helmholz Alliance on Systems Biology network SB-Cancer and the German Federal Ministry of Education and Research within the NGFNplus grants IG Cellular Systems Genomics and IG Prostate-Cancer.

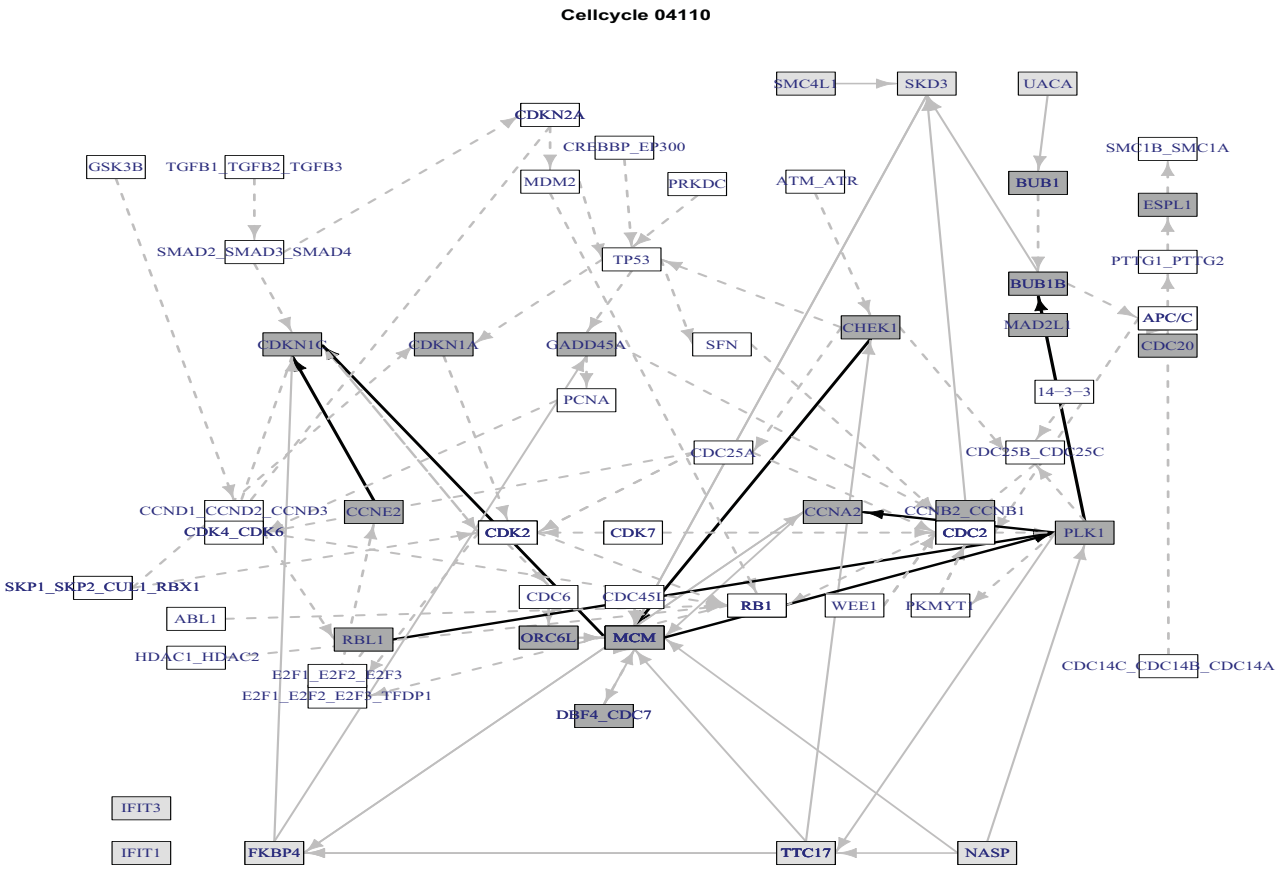


Figure 1: DBN network merged with the Cell Cycle KEGG network *hsa:04110*. White nodes are nodes only present in the KEGG network, dark grey nodes are present in both the DBN and KEGG network and light grey nodes are only in the DBN net. Dashed grey edges are found in KEGG, solid grey edges are predicted by the DBN and solid black edges are predictions that could be verified by the Ingenuity®literature network.

References

Affara, M., B. Dunmore, C. Savoie, S. Imoto, Y. Tamada, H. Araki, D. S. Charnock-Jones, S. Miyano, and C. Print (2007, Aug). Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? *Philos Trans R Soc Lond B Biol Sci 362*(1484), 1469–1487.

Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo (2004). Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics 20*(4), 578–580.

Beissbarth, T. (2006). Interpreting experimental results using gene ontologies. *Methods Enzymol 411*, 340–352.

Beissbarth, T. and T. P. Speed (2004). Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics 20*(9), 1464–1465.

Benjamini, J. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society 57*, 289–300.

Fröhlich, H., M. Fellmann, H. Sültmann, A. Poustka, and T. Beißbarth (2008). Predicting pathway membership via domain signatures.

Hahne, F., A. Mehrle, D. Arlt, A. Poustka, S. Wiemann, and T. Beissbarth (2008). Extending pathways based on gene lists using interpro domain signatures. *BMC Bioinformatics 9*, 3.

Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18 Suppl 1*(4), S96–104.

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi (2008). Kegg for linking genomes to life and the environment. *Nucleic Acids Res. 36*, D480 – D484.

Lebre, S. (2007). Inferring dynamic genetic networks with low order independencies.

Mulder, N. J., R. Apweiler, and T. K. Attwood et.al. (2008). New developments in the interpro database. *Nucleic Acids Res. 35*, D224 – D228.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol 3*(40), Article3.

Systems, I. Ingenuity pathways analysis, www.ingenuity.com.

The Gene Ontology Consortium (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Research 32*, D258–D261.

Modeling of microarray time-course data with dynamic Bayesian networks and within-time-point interaction

Brian Godsey[1] and Peter Sykacek[1]

[1] Chair of Bioinformatics, Department of Biotechnology,
Boku University, A-1190 Vienna, Austria
contact: brian.godsey@boku.ac.at

October 22, 2008

1 Introduction

State space models are increasingly popular in the area of microarray time-course analysis. These models represent large-dimensional data and parameters as points (states) in large-dimensional spaces, for which interaction and dependencies are appropriately defined. These models, particularly in the case of dynamic Bayesian networks (DBNs), are well-suited for capturing the behavior of data in time-course experiments and drawing conclusions about genetic interactions.[1, 4, 6]

Dynamic Bayesian networks are simply standard Bayesian networks combined with a time dimension. In these networks, there is one element (node) in the model for each gene at each time point. Often, this means hundreds or thousands of nodes in the network. A standard assumption, to ensure the definability of the model, is that all directed edges lead from nodes in one time point to nodes in the next. In other words, while standard (static) Bayesian networks with nodes representing genes run the risk of containing directed cycles as we attempt to infer which edges are most likely, DBNs circumvent this problem by allowing edges only from one time point to the next. For example, data points in time point 1, influence only the data points in time point 2, which in turn only influence time point 3, and so on. This has the desirable effect that, by definition, no directed cycles are introduced into the network graph, and furthermore that the element of time is introduced so that the past influences the present, and the present influences the future, with respect to a particular time point.[6]

However, these models are ignoring that some genes may be influenced by other genes within the same time point, and not just the genes from the previous time point. For instance, the time course data presented in Affara, *et al.*[1]

contains samples from time points 0, 0.5, 1.5, 3, 6, 9, 12, and 24 hours in an experiment monitoring human endothelial cells during the process of apoptosis. The smallest interval between samples is thirty minutes, while the average interval between samples is over three hours. Some gene-related reactions have been shown to operate on a shorter time-scale than that.[5, 2] While time-series experiments are generally designed considering the appropriate time-scale for the given organism and cell type, there is a significant chance that some RNA activity occurs at a rate fast enough to be unobservable in the design. That is to say that, if some gene can be activated or deactivated over a period of five or ten minutes, a common state space model such as in Kim *et al.*,[6] or Beal *et al.*[4], may not be able to detect that genetic interaction, because such a quick response may occur entirely between time points, or the expression changes of the two genes involved may straddle a time point, thus rendering their interaction undetectable. In general, undersampling can inhibit and otherwise cause problems with gene interaction inference.[3]

Assume, for instance in the Affara data set, that two genes have an interaction which operates on a time scale of approximately ten minutes. In other words, when the expression level of gene A changes, for whatever reason, gene B's expression level reacts accordingly within ten minutes. If gene A's expression increases just before time point t, and decreases just after, the full response of gene B will likely occur between time points t and $t+1$, thus rendering the interaction imperceptible in this case. Gene interactions operating on a short time-scale are more likely to be detectable within each of the time points, in a measure related to correlation, when compared to detection among relatively sparse sample points. However, one problem that arises when attempting to detect interactions within time points is that it is often very difficult, if not impossible, to determine which gene regulates and which gene is regulated.

Thus, we build upon existing methods of inferring interactions in a DBN setting, and include further measures to acknowledge that some genes change expression more quickly than the data allow us to witness.

2 Methods

In order to infer fast gene interaction, we used a k-means clustering algorithm to create groups of genes that have very similar expression patterns across the time points. Members of a given cluster likely either (1) interact with one another on a short time-scale, or (2) have a common regulator. In the latter case, if the regulator operates on a short time-scale, it would be in the same cluster, and if operating on a longer time scale, it would be in a different cluster.

Clusters are generated using a standard k-means clustering algorithm, with the appropriate number of clusters selected according to Akaike's information criterion, and the assumption that cluster membership likelihood functions are Gaussian with mean at the cluster center. Thus, an estimate of the likelihood for the current clustering arrangement can be calculated, and we can be certain that, for the given data, algorithm reaches an optimal solution. Before

clustering, the time profiles of the genes are normalized to the same variance. This acknowledges and corrects for the case that two genes, even if perfectly correlated, would not end up in the same cluster if one was much more highly expressed than the other.

After clustering, we used the variational Bayesian (VB) algorithm for DBN models with hidden states presented in Beal, *et al.*[4] to estimate the interaction matrix between the cluster centers. This algorithm infers dependencies in a linear DBN, as in Kim, *et al.*, [6], but with the further addition of hidden states. That is, Beal's model includes at each time point both observable and unobservable data points. The observable nodes of course represent gene expression measurements, usually from microarrays, whereas the hidden or unobservable nodes can represent any value in this system for which we don't have a measurement, such as genes not measured in the experiment or perhaps some sort of non-RNA gene regulator. In Beal's design, these hidden states are allowed to have an influence on genes in the same time point as well as on genes and hidden states in the following time point. In this way, directed cycles are still avoided, while allowing some states (the hidden states) to affect other states (observed states) within the same time point. Even though the hidden states interact with observed states within each given time point, direct gene-gene interaction between observable genes within a given time point is not considered. Thus, we utilize Beal's method, but expand it to consider interaction on a time-scale shorter than the sampling intervals, by utilizing clustering.

The results from the variational Bayesian model estimation should indicate which of the cluster centers interact with each other across the time points, thus adding to our knowledge of gene interactions taking place on a longer time scale than the sampling intervals. If we wish to see which interactions take place on shorter time scales, we can look inside the clusters.

A possible concern would be that gene-gene interaction may not be detected due to the genes being in the same cluster, and thus their respective nodes in the state space model would have no chance of being connected when fitting the model. In practice, this is no problem, because membership in the same cluster implies a high correlation between the genes' time profiles. A high time-offset correlation (as in, the time profile of gene A correlates with the profile of gene B offset by one time point), which would be present if the genes did, in fact, have an interaction as tested by our DBN, contradicts a direct correlation. Therefore, the strongest interactions are found by fitting the DBN in the case of longer time-scale interactions, and by looking within clusters for genes related to each other on shorter time scales.

3 Results and conclusion

The optimal number of clusters of the 18451 genes was 273, with the smallest cluster containing 10 genes and the largest containing 225.

Upon running the VB model-fitting algorithm, we found few significant individual interactions. In fact, only 3 interactions were discovered at a significance

level of $P < 0.05$, and 14 interactions at $P < 0.25$. While this is somewhat unfortunate, it is also to be expected. When fitting a model to discover the dynamics of 273 cluster centers, there are $273 * 272 = 74,256$ possible edges, meaning not only that we are trying to fit very large numbers of parameters at once, but also that there are many possible combinations of those parameters that could be equally likely. Of course, the three cluster-cluster interactions indicate that there are many gene-gene interactions, as all of the genes in a given cluster can have a joint influence on the genes of another cluster.

The three strongest interactions, those having significance $P < 0.05$, are found at edges $40 \rightarrow 83$, $61 \rightarrow 83$, and $128 \rightarrow 164$, (where the numbers are arbitrary cluster labels for the 273 clusters) with significance $P < 0.0005$, $P < 0.0010$, and $P < 0.0500$, respectively. None of the clusters 40, 61, or 128 have any parent clusters of any reasonable significance in the model. Thus, we have a strong reason to believe that the genes in clusters 40, 61, and 128 play important roles as the earliest and strongest-acting regulators of apoptosis. It would seem that cluster 83 plays a later role in apoptosis, activating after the main regulators. The accession numbers for clusters 40, 61, and 83 can be found in an addendum to this paper. Clearly, there are many genes in each cluster, and the list may be too long to be fully tested in apoptosis regulation. But, in the data set studied, the genes within each cluster are highly correlated with each other, and it is likely that the task of determining regulators *within* a given cluster is impossible, given the data set. An algorithm such as that used here would require more sampling points in the data collection to be able to distinguish between genes within the clusters.

As reported in Affara, *et al.*[1], the sequence known as GABARAP had the most number of children in a fitted state-space model. We found GABARAP to be in cluster 138, whose most significant child in our model does no better than $P < 0.40$. This does not necessarily contradict these previous results, as we merely suggest other possible significant interactions. On the other hand, our results agree with those in [1] in the lack of any significant parents for GABARAP (or its cluster).

In conclusion, we present a method of detecting regulatory genes as well as interaction between genes. The method, based on gene clustering and fitting a dynamic Bayesian model, leaves us with three highly significant interactions between gene groups. We propose that the genes listed as members of clusters 40, 61, and 128 are likely candidates for apoptosis regulators.

Acknowledgements

The Boku Bioinformatics group acknowledges support by the Vienna Science and Technology Fund (WWTF), Baxter AG, Austrian Research Centres (ARC) Seibersdorf, and Austrian Centre of Biopharmaceutical Technology (ACBT).

References

[1] M. Affara, B. Dunmore, C. Savoie, S. Imoto, Y. Tamada, H. Araki, D. S. Charnock-Jones, S. Miyano, and C. Print. Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? *Phil. Trans. R. Soc. B*, 2007. Published online.

[2] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16), 2004.

[3] S.D. Bay, L. Chrisman, A. Pohorille, and J. Shrager. Temporal aggregation bias and inference of causal regulatory networks. *Journal of Computational Biology*, 11(5), 2004.

[4] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356, 2004.

[5] S. Cokus, S. Rose, D. Haynor, N. Grønbech-Jensen, and M. Pellegrini. Modelling the network of cell cycle transcription factors in the yeast *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7, 2006.

[6] S. Y. Kim, S. Imoto, and S. Miyano. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings in Bioinformatics*, 4(3):228–235, September 2003.

4 Gene lists

4.1 Accession numbers for cluster 40

1012837.1, AB007974, AB014514, AB040878, AF054994, AF086381, AF132811, AK057549, AI674974, AJ133439, AK026158, AL049455, AL050152, AL122109, AL137259, BC011693, NM_000292, NM_000297, NM_000433, NM_000633, NM_000693, NM_001231, NM_001466, NM_001872, NM_002216, NM_003635, NM_003970, NM_004115, NM_004352, NM_004988, NM_006157, NM_006361, NM_006565, NM_006856, NM_006875, NM_007168, NM_007237, NM_012240, NM_014329, NM_014799, NM_015839, NM_016509, NM_016546, NM_017589, NM_017632, NM_017772, NM_019849, NM_030820, U07561

4.2 Accession numbers for cluster 61

BG717745, AI246523, AB014542, AF232772, AK001976, AK024432, AL050143, AL512713, NM_000236, NM_000272, NM_000275, NM_000296, NM_000855, NM_000898, NM_001407, NM_001637, NM_001741, NM_002298, NM_002664, NM_003640, NM_004207, NM_004737, NM_005919, NM_006225, NM_006757, NM_012260, NM_014037, NM_014400, NM_014860, NM_014871, NM_016589, NM_017451, NM_017697, NM_018335, NM_019063, NM_024500, U79277

4.3 Accession numbers for cluster 83

AJ249369, L38290, 1398420.5, AJ237736, BC018063, 221907.1, AL833218, NM_003385, NM_001083, AB054575, NM_018938, BF692587, BC029526, NM_002469, AB020676, AF010236, AF131756, AF131784, AI524085, AK000681, AK001442, AI923217, M31774, NM_000256, NM_000339, NM_000549, NM_000740, NM_001778, NM_003293, NM_005218, NM_005291, NM_005577, NM_006789, NM_012211, NM_014516, NM_015596, NM_017767, NM_020957, NM_022440, NM_024492, X52001, NM_005366,

Inference of Key Transcriptional Regulators in Endothelial Cell Apoptosis using Bayesian State Space Models

Claudia Rangel-Escareño, Irma Aguilar-Delfin
National Institute of Genomic Medicine, México City
Claremont Graduate University, Claremont, USA
Email: crangel@inmegen.gob.mx, iaguilar@inmegen.gob.mx

David L. Wild
Systems Biology Centre, University of Warwick, Coventry, UK
Keck Graduate Institute, Claremont, USA
Email: D.L.Wild@warwick.ac.uk

1 The proposed analytical objective

The contest dataset describes with the response of human vascular endothelial cells (HUVEC) to serum withdrawal, triggering apoptosis. The dataset is typical in that a complex biological phenomenon is probed by a timecourse with only a few measurements. It provides the classical challenge to microarray data analysis of extracting insight in a data space of very uneven dimensionalities. The challenge in this case is to identify candidate regulators.

2 A brief summary of the analytical effort

We have adopted the strategy for 'top-down' regulatory network identification using the input-driven variational Bayesian state space modelling approach of Beal et al. (1) that we have successfully utilized in other projects. State space models (SSMs) have a number of features which make them attractive for modelling gene expression time series data. They assume the existence of a set of hidden state variables, from which noisy continuous measurements can be made, and which evolve with Markovian dynamics. In our application, the noisy measurements are the observed gene expression levels at each time point, and a key innovation of our method is that we assume that the hidden variables are modelling effects which cannot be measured in a gene expression profiling experiment. The effects of genes which have not been included on the microarray, levels of regulatory proteins, the effects of mRNA and protein degradation are examples of such hidden variables.

3 Data Analysis

3.1 Data Normalization

Normalization is performed to reduce the systematic variation inherent to sample-array processing. There are several normalization methods and so far none of them is considered to be a gold standard. In general, the decision about which method to use should be based on reducing noise without affecting the biological signal to the point of either losing a target or including too many false positives. Previous work using CodeLink Bioarrays on a time course experiment (2) showed that median normalization (the manufacturer-recommended approach for this this platform) does not perform as well as other approaches. In this study it is shown that cyclic loess (pair-wise loess) performs better at reducing variability more effectively and consistently than median normalization. To check these conclusions, we decided to analyze the performance of two methods on the HUVEC time course data - normalization with the median and loess. This time course data consists of 8 time points and 3 pools of cells from 10 individuals, making a total of 20469 targets. Figure 1 shows density and box plots of the raw (top), the median (middle) and the loess (bottom) normalized data. Looking at the raw

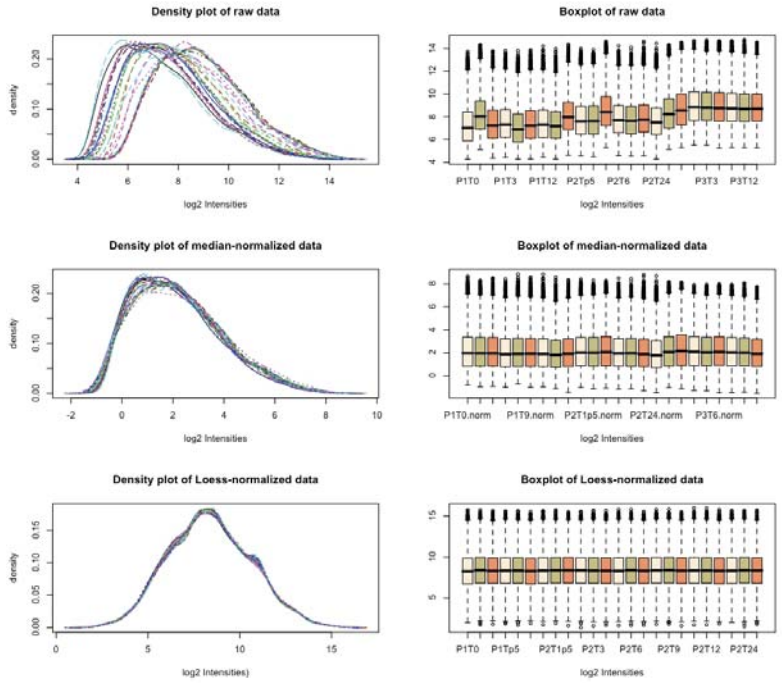


Figure 1: Density curves and boxplots for raw data (top), median-normalized data (middle) and loess-normalized data (bottom)

intensities, it is easy to see that a normalization is required to make the arrays comparable and also standardize their dynamic ranges. When we look at the effect of the median normalization it seems to be doing a good job. However, if we analyze another perspective of the same data – the boxplots (middle right)– we can see that the mean lines are not as well aligned as they are when using loess. Also, by using loess, the intensities remain within a similar dynamic range. That is, the raw intensity values are roughly within $(4, 14)$, which is a fairly common case. Using the median normalization the intensities fall within $(-1, 8)$, and negative intensity values are physically unrealistic. However, this is easily resolved by shifting the median to the right. However, the loess normalization approach extends the range on both the lower and upper limits by roughly one unit and is seen to produce an approximately normal intensity distribution. This analysis is based on the consensus information from the three pools. In our subsequent analysis, therefore, the loess normalized data were used. After removing the probes flagged with error labels and matching the good ones in all 24 arrays, we ended up with 9848 probes.

3.2 Differential Expression Analysis

In this analysis, we are actually concerned with differential expression *over a time course*. Standard methods used to detect differential expression across two or more independent sample groups may not be appropriate to detect differential expression in time series data, since they do not address the fact that microarray time course samples may be correlated. Tai and Speed (3) have described a method for ranking and selecting genes from replicated microarray time course data with one or more biological conditions, based on a multivariate empirical Bayes log-odds score or Hotelling T^2 statistic. After normalization, this method was applied to produce a ranking of differentially expressed genes by order of the Hotelling T^2 statistic, using the *timecourse* package implemented in R/Bioconductor. Modulated expression across the time course is clearly visible in the top ranked genes, and undetectable in the bottom ranked genes.

3.3 Network Inference

Networks were then inferred from the microarray time series data using the variational state space modelling (VBSSM) approach of Beal et al. (1). A key feature of our approach to network inference is that it uses a fully Bayesian analysis, which avoids overfitting and provides error bars on all model parameters. In practice, a Bayesian learning scheme infers distributions over all the parameters and makes modelling predictions by

taking into account all possible parameter settings. In doing so we penalise models with too many parameters, embodying an automatic Occam's Razor effect. First, the question of how many hidden factors should be used to account for the dependencies in the observed data is answered by employing Bayesian model selection, a well-founded principle used in machine learning and statistics to choose between models of differing complexities. The VBSSM algorithm calculates a lower bound on the marginal likelihood (Bayesian *evidence*), $\mathcal{F}$, and plotting this against the dimensionality of the hidden state space, k, allows us to select an optimal model for the data set – that which maximizes the marginal likelihood (Figure 2). The variational Bayesian model also provides us with posterior distributions for the model parameters from which a connectivity matrix which describes all gene-gene interactions *over successive time points* may be derived. Details of this procedure are described in (1). We consider an element of this matrix as providing evidence for a candidate gene-gene interaction if the element's posterior distribution is positioned *significantly far from the zero point* of no influence. Significance in this scenario corresponds to the zero point being more than n standard deviations from the posterior mean for that entry. Since these distributions are Gaussian, and may lie above or below the zero point (corresponding to positive or negative regulation), we can use the standard Z-statistic for normally distributed variables to threshold the connectivity matrix at any desired level of statistical significance. The output from this procedure is a directed graph in which arrows are drawn *from* a gene expression variable at a given time t, *to* another gene variable whose expression it influences at the next time point, $t+1$ (see Figure 3).

It should be noted that the method we describe is not a genome-wide modelling excercise. Previous authors have already pointed out that there is a theoretical limit to our ability to infer gene regulatory networks from data that describe the dynamics of cell response (6; 7; 8; 9). This limit is related to the amount of data that needs to be acquired to avoid overfitting of model parameter estimates, particularly when maximum likelihood methods are used. Our earlier studies with both synthetic and experimental data (4; 5) indicate that, with the number of time points and biological replicates in the CAMDA data set, we would be able to model effectively the inter-relationships of around 50 genes in one computational experiment. We have therefore chosen to model the top 50 genes as ranked by the Hotelling T^2 statistic.

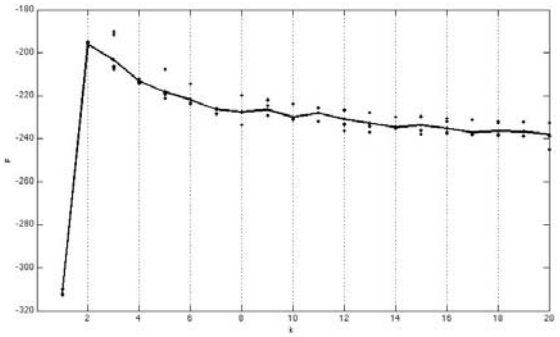


Figure 2: $\mathcal{F}$ versus k plot to select the optimal state space dimensionality - for this data set $k_{max} = 2$

4 Interpretation of results and presentation of discoveries in a biological context.

In terms of the challenge, *candidate regulators* are identified as those genes predicted to be major hubs in our inferred network (Figure 3). A number of these appear to be biologically plausible in the context of the processes triggered in endothelial cells under growth factor deprivation conditions:

- CDKN1C encodes the protein known as p57/Kip2, one of the cyclin-dependent kinase inhibitor proteins that function as negative regulators of the cell-cycle (14) and directly promote the the intrinsic apoptotic pathway (15). Smad-mediated transcription (SMAD1 is down-regulated by MTX1 in the network) has been shown to be involved in the induction of p57/Kip2 proteolysis (10). Although we would not expect to observe post-translational regulation directly in our model, such information could be included as a Bayesian prior in future cycles of modelling. One possible mechanism of SMAD1 downregulation could be via TGF-β signaling through the receptor ALK1, which is known to be expressed constitutively in endothelial cells (12). TGF-β signaling occurs under conditions of serum deprivation in endothelial cells, which, in turn, attenuates apoptotic death (13).

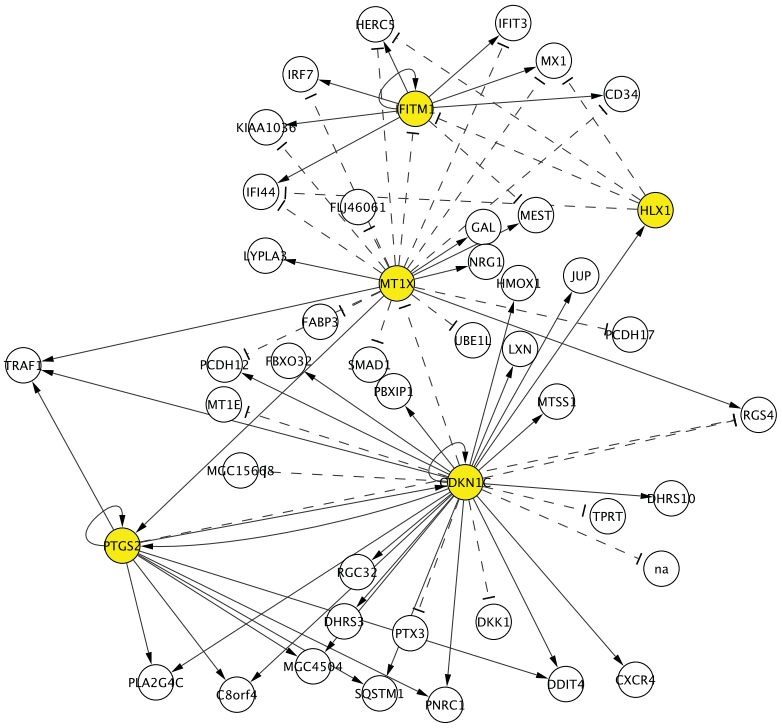


Figure 3: Gene regulatory network for the top 50 genes as ranked by the Hotelling T^2 statistic, at a confidence level of $>$ 99% (Z=3). Arrows represent gene-to-gene expression influences across consecutive time points: solid arrows represent up-regulation, dotted T's represent down-regulation.

- MT1X, another major hub, which is predicted to play a largely inhibitory role in the network, encodes a metallothionein, a family of low molecular weight proteins with a high affinity for divalent metals. Metallothionein has been shown to have a protective role in apoptosis, specifically by controlling the cellular zinc ion levels: the proper intracellular Zn2+ level maintains the fragmentation of DNA associated with caspase-3 activity (16).
- PTGS2, better known as COX-2, is an enzyme responsible for the synthesis of the signaling lipids known as prostaglandins. Stressful stimuli in endothelial cells have been shown to induce COX-2 expression and activity in the form of PGE2 production, which in turn triggers the caspase-3 activity that facilitates apoptosis (17)
- HLX1 is a homeobox gene transcription factor that has shown to regulate vascular development in embryonic and adult tissues (21). Its role in endothelial cell apoptosis is unexplored.
- IFITM1, interferon-induced transmembrane protein 1, has been shown to influence proliferation in response to the cytokine IFN-gamma. In our network we observe that it is predicted to play the role of an activator of other interferon-related genes. In hepatocytes, IFITM1 overexpression negatively regulates cell growth, whereas its suppression enhances it. Furthermore, IFITM1 inhibits the activity of extracellular signal-regulated kinase (ERK) and enhances the transcriptional activity of the tumor supressor gene p53 (22). These findings constitute a likely lead for the involvement of IFITM1 in the regulatory network, although there is currently no published data on IFITM1 on either endothelial cells or in the context of apoptosis.

5 Discussion

Interestingly, Hirose et al. (18) have recently published a report of an analysis of the same data set using a canonical state-space model (without input-driven feedback) and maximum likelihood parameter estimation. They inferred a network which shows TRAF1 as a major network hub, which up-regulates CDKN1C. TRAF1 encodes one of the TNF-receptor-associated factors, cytoplasmic adaptor proteins

that mediate cytokine signaling. The regulatory association predicted by Hirose et al. is in direct contradiction to our model, which predicts TRAF1 to be up-regulated by CDKN1C. These two models, therefore, represent contradictory, but *experimentally testable* hypotheses, which could be tested by gene silencing experiments on both CDKN1C and TRAF1. We suggest that such experiments would make an ideal follow-up to the CAMDA challenge.

References

[1] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild. **(2005)**. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21:349–356, 2005.

[2] Wu W., Dave N., Tseng G.C., Richards T., Xing E.P. and Kaminski N. **(2005)**. Comparison of normalization methods for CodeLink Bioarray data. *BMC Bioinformatics*, 6:309.

[3] Y. C. Tai and T. P. Speed. **(2006)**. A multivariate empirical bayes statistic for replicated microarray time course data. *Ann. Statist.*, 34, 2387–2412.

[4] C. Rangel, J. Angus, Z. Ghahramani, and D. L. Wild. Modeling genetic regulatory networks using gene expression profiling and state space models. In D. Husmeier, S. Roberts, and R. Dybowski, editors, *Probabilistic Modelling in Bioinformatics and Medical Informatics*, Springer Verlag, 2004.

[5] M.J. Beal, J. Li, Zoubin Ghahramani, and D.L. Wild. Reconstructing transcriptional networks using gene expression profiling and Bayesian state space models. In S. Choi, editor, *Introduction to Systems Biology*, pages 104–113. Humana Press, 2007.

[6] V. A. Smith, E. D. Jarvis, and A. J. Hartemink. Evaluating functional network influence using simulations of complex biological systems. *Bioinformatics*, 18(1):S216–S224, 2002.

[7] M. K. Yeung, J. Tegner, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci.*, 99:6163–6168, 2002.

[8] D. E. Zak, F. J. Doyle, G. E. Gonye, and J. S. Schwaber. Simulation studies for the identification of genetic networks from cdna array and regulatory activity data. In *Proceedings of the 2nd International Conference on Systems Biology*, pages 231–238. Omipress, Madison, WI, 2001.

[9] D. E. Zak, G. E. Gonye, J. S. Schwaber, and F. J. Doyle 3rd. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. *Genome Research*, 13:2396–2405, 2003.

[10] Nishimori S., Tanaka Y., Chiba T., Fujii M., Imamura T., Miyazono K., Ogasawara T., Kawaguchi H., Igarashi T., Fujita T., Tanaka K. and Toyoshima H. **(2001)**. Smad-mediated Transcription Is Required for Transforming Growth Factor–1-induced $p57^{Kip2}$ Proteolysis in Osteoblastic Cells *J. of Biological Chemistry* Vol. 276, No. 14, pp. 10700-10705

[11] Mori Y, Chen SJ, Varga J. **(2001)** Modulation of endogenous Smad expression in normal skin fibroblasts by transforming growth factor-beta. *Exp Cell Res* Aug 1;258(2):374-83.

[12] Ota T, Fujii M, Sugizaki T, Ishii M, Miyazawa K, Aburatani H, Miyazono K. **(2002)** Targets of transcriptional regulation by two distinct type I receptors for transforming growth factor-beta in human umbilical vein endothelial cells. *J Cell Physiol* Dec;193(3):299-318.

[13] Solovyan VT and Keski-Oja J **(2005)**. Apoptosis of human endothelial cells is accompanied by proteolytic processing of latent TGF-β binding proteins and activation of TGF-β. *Cell Death and Differentiation* 12, 815-826

[14] Urano T., Yashiroda H., Muraoka M., Tanaka K., Hosoi T., Inoue S., Ouchi Y., Tanaka K. and Toyoshima H. **(1999)**. $p57^{Kip2}$ Is Degraded through the Proteasome in Osteoblasts Stimulated to Proliferation by Transforming Growth Factor β 1. *J. of Biological Chemistry* Vol. 274, No. 18. pp.12197–12200

[15] Vlachos P., Nyman U., Hajji N., Joseph B. **(2007)**. The cell cycle inhibitor p57(Kip2) promotes cell death via the mitochondrial apoptotic pathway. *Cell Death Differ.* 14(8):1497-507

[16] Dutsch-Wicherek M, Sikora J, Tomaszewska R. **(2008)** The possible biological role of metallothionein in apoptosis. *Front Biosci.* May 1;13:4029-38.

[17] Sheu ML, Ho FM, Yang RS, Chao KF, Lin WW, Lin-Shiau SY, Liu SH. **(2005)** High glucose induces human endothelial cell apoptosis through a phosphoinositide 3-kinase-regulated cyclooxygenase-2 pathway. *Arterioscler Thromb Vasc Biol.* Mar;25(3):539-45.

[18] Hirose O , Yoshida R, Imoto S , Yamaguchi R , Higuchi T, Charnock-Jones DS, Print C, Satoru Miyano S **(2008)** Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models *Bioinformatics* 24: 932–942.

[19] Zirlik A, Bavendiek U, Libby P, MacFarlane L, Gerdes N, Jagielska J, Ernst S, Aikawa M, Nakano H, Tsitsikov E, Schönbeck U. **(2007)** TRAF-1, -2, -3, -5, and -6 are induced in atherosclerotic plaques and differentially mediate proinflammatory functions of CD40L in endothelial cells. *Arterioscler Thromb Vasc Biol.* May;27(5):1101-7.

[20] Wang Q, Dziarski R, Kirschning CJ, Muzio M, Gupta D. **(2001)** Micrococci and peptidoglycan activate TLR2→MyD88→IRAK→TRAF→NIK→IKK→NF-kappaB signal transduction pathway that induces transcription of interleukin-8. *Infect Immun.* Apr;69(4):2270-6.

[21] Murthi P, So M, Gude NM, Doherty VL, Brennecke SP, Kalionis B. **(2007)** Homeobox genes are differentially expressed in macrovascular human umbilical vein endothelial cells and microvascular placental endothelial cells. *Placenta*. Feb-Mar;28(2-3):219-23.

[22] Yang G, Xu Y, Chen X, Hu G. **(2007)** IFITM1 plays an essential role in the antiproliferative action of interferon-gamma. *Oncogene* Jan 25;26(4):594-603.

POSTERS

PARALLEL ANALYSES OF ARCHIVED MICROARRAY DATA SETS PROVIDE NEW BIOLOGICAL INFORMATIONS

M. PIERRE[1], A. GAIGNEAUX[1], F. BERGER[1], B. DEHERTOGH[1], E. BAREKE[1], C. MICHIELS[2] AND E. DEPIEREUX[1]

([1] Unité de Recherche en Biologie Moléculaire, Facultés Universitaires Notre Dame de la Paix, Namur, Belgium and [2] Unité de Recherche en Biologie Cellulaire, Facultés Universitaires Notre Dame de la Paix, Namur, Belgium)

Background : Today, DNA microarrays have become common tools in many laboratories of molecular biology and medecine because they enable the researchers to measure the expression of an entire genome in a single experience. However lots of microarray data sets have been published without being fully exploited.
Growing tumors are caracterized by hypoxia areas because pre-existing vasculature is outgrown and because new vasculature is abnormal. It is now accepted that hypoxia selects cancer cells able to survive and to migrate to distant organs because they exhibit a specific transcriptome, and thus a specific proteome.
In an effort to find new genes involved in metastasis and expressed in hypoxia, we have re-analysed 22 Affymetrix data sets related to our subject of interest. We attempted to get new reliable genes by combining several original approaches.

Methods : The first step was to analyse each data set individually. We used alternative Chip File Definition (CDF). A CDF is a file developed by Affymetrix for each platform, that links several probes (probe set) to a given gene name. The probes representing the « gene » reflect sometimes the status of genomic databases several years ago. Since then genomic information, and the arguments to assign probes to probe sets have evolved and alternative CDFs have emerged. To pre-process the data (background correction, normalisation and summarization of signal values measured on a chip) we used GCRMA. And to process the data (statistical evaluation of the differential expression of genes between two conditions), we used the Window *t* test, a modified version of the classical Student *t* test.
Out of these individual analyses, we got volcano plots, graphs showing log10 of fold change on the X axis and –log10 of p values (statistical significance) on the Y axis for each probe set on the chip. To select genes of interest for biological tests out of these volcano plots, we used three different apprroaches.
The first one was to consider the most significant genes common to several data sets. We named this approach « the intersections ». The second approach was to select the most significant genes common to at least one « metastasis data set » and to at least one « hypoxia data set ». We named this approach « the union intersections ». The last approach was the meta-analysis of the data sets. Here we pre-processed and processed several data sets as one to increase the number of replicates and thus to gain statistical power.

Results : 33 intersections were designed with different biological relevance in function of which data sets were included. Table 1 presents the categories of intersections (column 1), the number of intersections (column 2), the mean number of data sets included (column 3) and the mean number of genes considered to get 50 genes (our limit for *in vitro* tests) common to the data sets. Intersections provide us 704 different genes.

Categories	Intersections	Data sets	Genes
Chip model	20	8	4892
Condition	5	13	16464
Tissue	7	3	1829
Total	1	20	22830

Table 1.

As for the intersections, 30 union intersections were designed. Table 2 presents the categories of union intersections (column 1), the number of union intersections (column 2), the mean number of metastasis data sets and hypoxia data sets included respectively (column 3) and the mean number of genes considered to get 50 genes common to the data sets. Union intersections provide us 245 different genes.

Categories	Union intersections	Data sets	Genes
Chip model	19	8 U 3	184
Condition	4	16 U 3	129
Tissue	7	3 U 3	297

Table 2.

For the meta-analysis, 14 meta-data sets were designed. All were pre-processed using GCRMA and processed with the Window *t* test. For each meta-data sets the 50 most significant genes were selected providing us 406 different genes.
Table 3 shows the number of genes common to several approaches. In total there are 117 genes that we consider of interest. Out of these 117 genes an uneglectable proportion is involved in metastasis confirming the validity of our approach. The future directions of this work are to learn more about the 117 genes in order to select 50 of them to test their expression *in vitro*.

Methods	Genes
Intersections / Union intersections	9
Intersections / Meta-analysis	107
Union intersections / Meta-analysis	3
Intersections / Union intersections / Meta-analysis	1

Table 3.

Acknowledgements
M. PIERRE is recipient of a FRIA fellowship.

Contact
M. PIERRE : michael.myst@hotmail.com

Non-specific hybridization scaling of microarray expression estimates – a physico-chemical approach for chip-to-chip normalization

Hans Binder[1*], Jan Brücker[1], Conrad Burden[2]

[1] Interdisciplinary Centre for Bioinformatics of Leipzig University, D-4107 Leipzig, Haertelstr. 16-18

[2] Centre for bioinformation Science, John Curtin School of Medical Research and Mathematical Sciences Institute, Australian National University, Canberra, A.C.T. 0200, Australia

* Corresponding author: E-mail: binder@izbi.uni-leipzig.de, fax: ++49-341-9716679

Key words: Expression analysis, microarray technology, oligonucleotide probes, background correction, DNA/RNA duplexes, base pair interactions, normalization rules

Abstract

Background: The problem of inferring accurate quantitative estimates of transcript abundances from gene expression microarray data is addressed. Particular attention is paid to correcting chip-to-chip variations arising mainly as a result of unwanted non-specific background hybridization, to give transcript abundances measured in a common scale. This study verifies and generalizes a model of the mutual dependence between non-specific background hybridization and the sensitivity of the specific signal using an approach based on the physical chemistry of surface hybridization.

Results: We have analyzed GeneChip oligonucleotide microarray data taken from a set of five benchmark experiments including dilution, Latin Square and "Golden Spike" designs. Our analysis concentrates on the important effect of changes in the unwanted non-specific background inherent in the technology due to changes in total RNA target concentration and/or composition. We find that incremental changes in non-specific background entail opposite sign incremental changes in the effective specific binding constant. This effect, which we refer to as the "up-down" effect, results from the subtle interplay of competing interactions between the probes and specific and non-specific targets at the chip surface and in bulk solution. We propose special rules for proper normalization of expression values considering the specifics of the up-down effect. Particularly for normalization one has to level the expression values of invariant expressed probes. Existing heuristic normalization techniques which do not exclude absent probes, level intensities instead of expression values and/or use low variance-criteria for identifying invariant sets of probes lead to biased results. Strengths and pitfalls of selected normalization methods are discussed. We also find that the extent of the up-down effect is modified if RNA targets are replaced by DNA targets, in that microarray sensitivity and specificity are improved via a decrease in non-specific background, which effectively amplifies specific binding.

Conclusions: The results emphasize the importance of physico-chemical approaches for improving heuristic normalization algorithms to proceed towards quantitative microarray data analysis.

“Hook”-calibration of GeneChip-microarrays

Hans Binder[1], Mario Fasold[1], Jan Brücker[1] and Stephan Preibisch[2]

[1] Interdisciplinary Centre for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany

[2] Max-Planck-Institute for Molecular Cell Biology and Genetics, D-01307 Dresden, Germany

Background: The improvement of microarray calibration methods is an essential prerequisite for quantitative expression analysis. This issue requires the formulation of an appropriate model describing the basic relationship between the probe intensity and the specific transcript concentration in a complex environment of competing interactions, the estimation of the magnitude these effects and their correction using the intensity information of a given chip and, finally the development of practicable algorithms which judge the quality of a particular hybridization and estimate the expression degree from the intensity values.

Results: We present the so-called hook-calibration method which co-processes the log-difference (delta) and –sum (sigma) of the perfect match (PM) and mismatch (MM) probe-intensities. The MM probes are utilized as an internal reference which is subjected to the same hybridization law as the PM, however with modified characteristics. After sequence-specific affinity correction the method fits the Langmuir-adsorption model to the smoothed delta-versus-sigma plot. The geometrical dimensions of this so-called hook-curve characterize the particular hybridization in terms of simple geometric parameters which provide information about the mean non-specific background intensity, the saturation value, the mean PM/MM-sensitivity gain and the fraction of absent probes. This graphical summary spans a metrics system for expression estimates in natural units such as the mean binding constants and the occupancy of the probe spots. The method is single-chip based, i.e. it separately uses the intensities for each selected chip.

Conclusions: The hook-method corrects the raw intensities for the non-specific background hybridization in a sequence-specific manner, for the potential saturation of the probe-spots with bound transcripts and for the sequence-specific binding of specific transcripts. The obtained chip characteristics in combination with the sensitivity corrected probe-intensity values provide expression estimates scaled in natural units which are given by the binding constants of the particular hybridization.

EMERALD – Enhancing Microarray Data Quality

Vidar Beisvag[1], Arne K. Sandvik[1] and Martin Kuiper[2,3]

[1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology.

[2]VIB, Department of Plant Systems Biology, University of Gent

[3]Department of Biology, Norwegian University of Science and technology.

Contact information: vidar.beisvag@ntnu.no

EMERALD is a European Coordination Action to develop measures to improve microarray data quality.

Microarray-based functional genomics technology suffers from a lack of standards. We are establishing and disseminate quality metrics, microarray standards and best laboratory practices throughout the European microarray community. We have developed specific quality metrics software (http://www.microarray-quality.org/quality_metrics.html), and we develop a Normalisation and Transformation ontology (http://www.microarray-quality.org/ontology_work.html) to allow for a structured recording of data pre-processing. The quality metrics software is being used to select high quality microarray (compendium) data from public databases. We expect that such quality-selected data will provide a valuable source for systems biology approaches like network inference and reverse engineering.

We hope to initiate European efforts to assess the merits of hybridisation standards for QC, and launch procedures to certify selected standards as European Reference Material. We are uniting the different microarray technology stakeholders in a series of topical workshops to address the development and implementation of QA/QC in research, service, diagnostics, data pre-processing and archiving, computational datamining, new technology development and its exploitation. We strive toward a wide community acceptance of 'best practices'. A web portal at EBI (http://www.microarray-quality.org/index.html) presents a network of contacts, and will serve to disseminate protocols, data, the use of control material, etc. We will assist individual microarray users in their transfer to such common practices. The results and experiences from transcriptome microarray QA/QC will create a cornerstone for a systems biology based life science, and cross-fertilise and advance the maturation process of emerging applications of microarray technology.

One Chip per condition: comparison of analysis methods for Affymetrix chips.

Anthoula GAIGNEAUX, Michael PIERRE , Benoît DE HERTOGH, Fabrice BERGER, Eric BAREKE, Eric DEPIEREUX

Molecular Biology Research Unit,University of Namur, Belgium

Mail to: anthoula.gaigneaux@fundp.ac.be

Because of the relative high price of microarrays chips, microarrays experiments are often performed with a low number of replicates. Several methods were developped to handle this limitation, most of them taking into account the high number of probes on a chip to otbain a better estimation of the variance by moderating the error accross probes. Beside these low replicate experiments, there is a demand for methods allowing the analysis of one chip per condition. These experiments are expected to serve as preliminary analysis for further chip or bench experiments, as no biological nor technical replicates are included.

Statistically, there is no way to compute a variance from one measure, meaning that the only way to analyse these experiments is to use of the FoldChange as a ranker. However, several methods exists to handle such experiments, taking advantage of the multiple probes per probesets present on chips like Affymetrix. These probes are usually summarized in one expression value per probeset, but can be used as is to allow a variance estimation.

The aim of this work was to compare several available methods (Foldchange, anova at probe-level [1], S-score [2], EBarrays [3], and PPLR [4]) handling one-chip design and to address some questions about the dataset- and the chip pair-dependence of results, and about the performance of such methods when used with replicate chips.

To assess method efficiencies, we used 2 spike-in experiments, chosen to have different characteristics, preprocessed and analyzed data, and derived values of True Positive Rate and False Discovery Rate for several alpha thresholds in order to build FDRoc curves.

The figure 1 outlines the main results of our study. It shows that results are very different with respect to the dataset used, which was expected from their very divergent characteristics. However method ranks are globally consistent between datasets. The comparison of curves obtained with several chip pairs shows that some methods are very chip-pair dependent, while datasets comprise only technical replicates. The figure 2 shows that, while not a statistical method, the simple computation of the Foldchange provides good results.

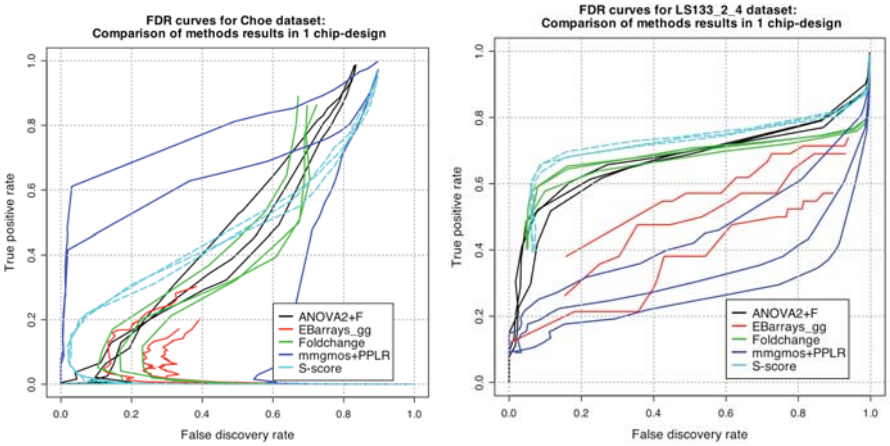


Figure 1 Choe and LatinSquare datasets: FDRoc curves for 1-chip design. For each dataset, 3 chip-pairs are considered.

In addition, we evaluate the efficiency of these methods when used with several replicates, and compare them to T and Baldi's regularized T tests [5]. We find curves ranks similars to those obtained with one chip, but methods generally do not outperformed Baldi's regularized T test.

Further work will include tests on simulated probe-level data.

Références

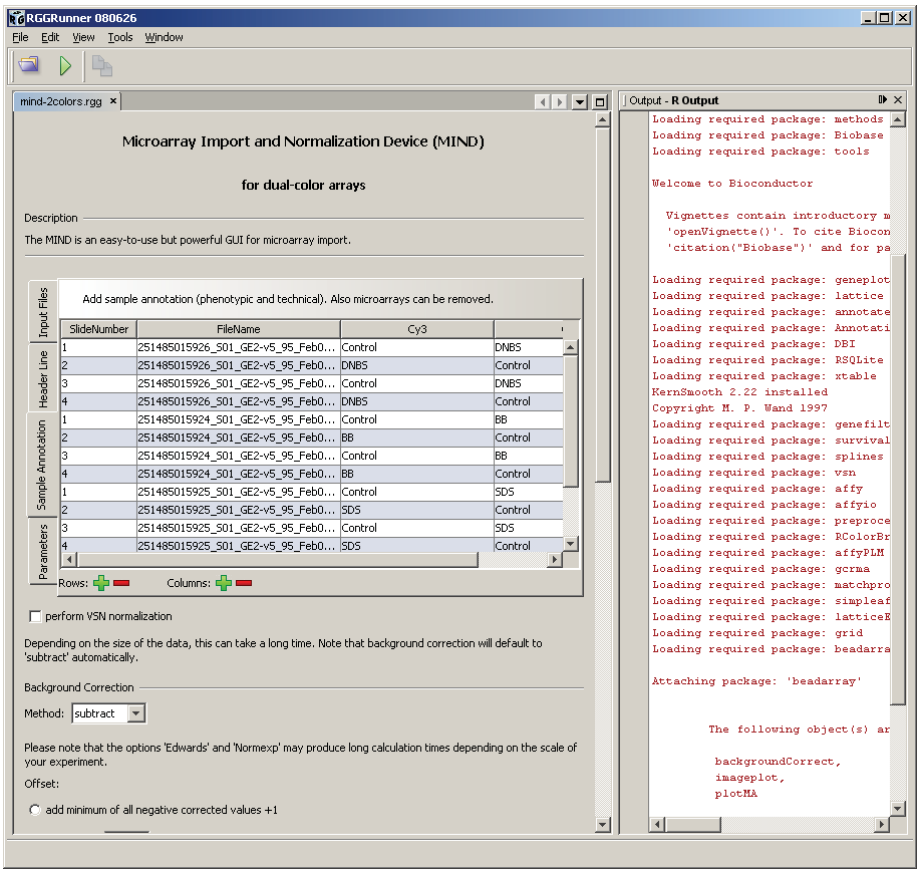
1. Lemieux, S., *Probe-level linear model fitting and mixture modeling results in high accuracy detection of differential gene expression.* BMC Bioinformatics, 2006. **7**: p. 391.
2. Kennedy, R.E., K.J. Archer, and M.F. Miles, *Empirical validation of the S-Score algorithm in the analysis of gene expression data.* BMC Bioinformatics, 2006. **7**: p. 154.
3. Kendziorski, C.M., et al., *On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.* Stat Med, 2003. **22**(24): p. 3899-914.
4. Liu, X., et al., *Probe-level measurement error improves accuracy in detecting differential gene expression.* Bioinformatics, 2006. **22**(17): p. 2107-13.
5. Baldi, P. and A.D. Long, *A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes.* Bioinformatics, 2001. **17**(6): p. 509-19.

RGG / MIND: On-the-fly GUI generation from R-scripts for Microarray analysis

Ilhami Visne[1], **Klemens Vierlinger**[1], Michael Srb[1], Christa Noehammer[1], Friedrich Leisch[2], Albert Kriegner[1]

[1]Austrian Research Centers GmbH -ARC, Life Sciences, Molecular Diagnostics, A-2444 Seibersdorf, Austria [2] Institut für Statistik Ludwig-Maximilians-Universität Muenchen

R is the leading open source statistic software with a vast number of analysis packages which are developed by a large user community (>100,000 users). However, the use of R requires programming skills. We have developed a GUI generator for R scripts based on a GUI definition language in XML. A GUI is generated by adding predefined GUI tags to the R script. User-GUI interactions are converted to R – Code, which replace the xml-tags in the R script. The project's aim is to provide R developers with a tool to make R based statistical computing available to a wider audience less familiar with script based programming. RGG can also benefit experienced R programmers by enabling to make regular changes to the code via a GUI interface (e.g. setting of an input file path via file chooser). RGG is a powerful tool to combine the advantages of script and GUI based computing while maintaining high flexibility. We showcase the utility of RGG using our Microarray Import and Normalisation Device (MIND), which imports Microarray data, lets the user choose the preprocessing procedures, produces a variety of QC plots and performs some explorative statistics. GUIs for inference statistics and machine learning are also available. The project further includes the development of a repository and documentation system for R-GUIs being developed by community.



A model based approach to probe design for high-performance microarray

German Gaston Leparc[1], Thomas Tüchler[1], Gerald Striedner[2], Karl Bayer[2], Peter Sykacek[1], Ivo Hofacker[3], and David Kreil[1]

[1]WWTF Chair of Bioinformatics, Universität für Bodenkultur Wien

[2]Institute of Applied Microbiology, Universität für Bodenkultur Wien

[3]Theoretical Biochemistry Group, Institute for Theoretical Chemistry, University of Vienna

DNA microarray technology has become a well-established, powerful high-throughput tool in biological research. Despite the huge success of microarray analyses, the interpretation of gene expression data remains a challenge. Many modern methods for microarray data analysis aim to detect biologically meaningful patterns or signatures in the data, and thus particularly rely on accurate measurements. Highly specific probes with uniform hybridization behaviour are therefore crucial for accurate quantitative modelling and further advancement of inference methods in microarray analysis. The main challenge in high-performance microarray design is the selection of highly specific oligonucleotide probes for all targeted genes of interest while, at the same time, maintaining thermodynamic probe uniformity. We introduce and describe a novel microarray design framework incorporating several advanced features for improved microarray performance and genome-scale designs. Our method improves on a number of aspects central to probe design: A quantitative model captures experimentally determined effects of probe placement along the target on labelling efficiency. In addition, the prediction of probe--target hybridization was improved by considering both probe and target structure. For efficiency, probe cross-hybridization predictions on genome scale usually exploit fast sequence-similarity based heuristics as filter before employing thermodynamic models. Of course, this brings a necessary trade-off between speed and sensitivity. There have, however, been no published studies so far examining the degree to which these heuristics may miss cross-hybridization targets. We present the results of a rigorous calibration of sequence based heuristics through sensitive thermodynamic calculations. The calibrated heuristic can then serve as conservative filter. Finally, we formulate a novel compound score, combining all probe features calculated in a principled way. This permits an objective selection of discriminative probe candidates while, at the same time, maintaining probe uniformity. By applying full global set optimization rather than a greedy search, our approach delivers maximally specific and unusually uniform probe sets. We demonstrate the performance of our novel probe design framework using results from different genomes.

Development of whole genome DNA microarrays for *Pichia pastoris*

Brigitte Gasser[1], **Alexandra Graf**[1], Martin Dragosits[1], Michael Sauer[2], Germán G. Leparc[3], Thomas Tüchler[3], David P. Kreil[3], Diethard Mattanovich[1,2]

[1]Institute of Applied Microbiology, Department of Biotechnology, University of Natural Resources and Applied Life Sciences Vienna, Austria

[2]School of Bioengineering, University of Applied Sciences FH Campus Wien, Vienna, Austria

[3]Vienna Science Chair of Bioinformatics, Department of Biotechnology, University of Natural Resources and Applied Life Sciences Vienna, Austria

DNA microarrays are regarded as a valuable tool for basic and applied research in microbiology. However, for many industrially important microorganisms, such as the yeast *Pichia pastoris*, the lack of commercially available microarrays still hampers physiological research. Presently, our understanding of protein secretion in *P. pastoris* is widely dependent on conclusions drawn from analogies to *Saccharomyces cerevisiae*. To close this gap for a yeast species employed for its high capacity to produce heterologous proteins, we developed whole genome DNA microarrays for *P. pastoris* on the basis of a commercially available draft genome. Up to date NCBI contains only 13 mRNAs and 158 genomic DNA/RNA sequences for *P. pastoris*, a commercially available gene prediction contained 5425 ORFs of which 3677 had an assigned function. This starting set of genes was complemented through de-novo gene finding and annotated using BLAST with a reciprocal best hit strategy. For the resulting set of candidate genes oligos were design using Agilent's online service (eArray) as well as a second program called TherMODO[1]. Both oligo designs were used in a stress response experiment and compared to results from a similar experiment with *S. cerevisiae* [2].

For the two oligo design strategies, we evaluated sensitivity to cross-hybridization as well as Tm (melting temperature) and delta G (free energy) distribution of probes. TherMODO designed probes for 15,035 sequences, of which only 665 were predicted as having cross-hybridization potential. Agilent's eArray designed probes for 15,150 sequences, of which 617 were marked for a cross-hybridization risk. TherMODO proofed to be more sensitive, filtering out sequences that were too short or had a certain amount of N nucleotides as well as finding more sequences that contained a potential cross-hyb risk. The distributions of DG and T_m of both designs as shown in figure 1 on the poster clearly that TherMODO designed probes are more uniform with respect to the Gibbs free energy DG and melting temperature, indicating a superior hybridization performance.

The microarrays made it for the first time possible to study genome-wide regulation in the important protein production host *P. pastoris*, giving novel insights into UPR regulation patterns [3]. The differences observed between *P. pastoris* and *S. cerevisiae* once again underline the importance of DNA microarrays for industrial production strains, instead of drawing conclusions from model organisms. Overexpression of *HAC1*, the most direct control for UPR genes, resulted in significant new understanding of this important regulatory pathway in *P. pastoris*, and generally in yeasts.

[1] Leparc et al. 2008. Nucleic Acid Research.
[2] Travers et al. 2000. Cell 101(3): 249-258.
[3] Graf et al. 2008. BMC Genomics 9:390.

PARTICIPANTS

Jose Manuel	Arteaga	jmarte@essex.ac.uk	University of Essex, U.K.
Anaïs	Bardet	anais.bardet@boku.ac.at	Boku University Vienna, Austria
Tim	Beissbarth	tim.beissbarth@googlemail.com	University of Göttingen, Germany
Vidar	Beisvåg	vidar.beisvag@ntnu.no	NTNU, Norway
Ralph	Beneke	ralph.beneke@tecan.com	TECAN, Austria
Hans	Binder	binder@rz.uni-leipzig.de	IZBI - University of Leipzig, Germany
Eva	Budinská	budinska@iba.muni.cz	IBA - Masaryk University, Czech Republic
Joaquin	Dopazo	jdopazo@cipf.es	CIPF, Spain
Mario	Fasold	fasold@gmail.com	IZBI - University of Leipzig, Germany
Livio	Finos	L.Finos@lumc.nl	LUMC, The Netherlands
Anthoula	Gaigneaux	anthoula.gaigneaux@fundp.ac.be	FUNDP, Belgium
Walter	Glaser	walter.glaser@univie.ac.at	Max F. Perutz Laboratories, Austria
Brian	Godsey	brian.godsey@boku.ac.at	Boku University Vienna, Austria
Jelle	Goeman	j.j.goeman@lumc.nl	Leiden University Medical Center, The Netherlands
Alexandra	Graf	alexandra.graf@boku.ac.at	Boku University Vienna, Austria
Max	Kauer	maximilianotto@gmx.at	CCRI St. Anna Vienna, Austria

Florian	Klinglmueller	float@lefant.net	Medical University of Vienna, Austria
Peter	Konings	peter.konings@esat.kuleuven.be	K.U. Leuven ESAT-SISTA, Belgium
Martin	Kuiper	kuiper@nt.ntnu.no	NTNU, Norway
David	Kreil	david.kreil@boku.ac.at	Boku University Vienna, Austria
Pawel	Labaj	pawel.labaj@boku.ac.at	Boku University Vienna, Austria
German	Leparc	german.leparc@boku.ac.at	Boku University Vienna, Austria
Yong	Li	yong.li@zbsa.uni-freiburg.de	University of Freiburg, Germany
Walter	Liggett	walter.liggett@nist.gov	NIST, U.S.A.
Simon	Lin	s-lin2@northwestern.edu	Northwestern University, U.S.A.
James	Malone	malone@ebi.ac.uk	EBI, U.K.
Ulrike	Mückstein	ulrike.mueckstein@boku.ac.at	Boku University Vienna, Austria
Alena	Mysickova	lenarty@hotmail.com	Humboldt University, Germany
Ron	Peterson	ron.l.peterson@comcast.net	Novartis, U.S.A.
Michael	Pierre	michael.myst@hotmail.com	FUNDP, Belgium
Olivia	Sanchez-Graillet	osanch@essex.ac.uk	University of Essex, U.K.
Theresa	Scharl	theresa.scharl@boku.ac.at	Boku University Vienna, Austria
Eran	Segal	eran.segal@weizmann.ac.il	Weizmann Institute, Israel
Bernhard	Spangl	bernhard.spangl@boku.ac.at	Boku University Vienna, Austria
Ewa	Stocka	fkstocka@gmail.com	AGH, Poland

Peter	Sykacek	peter.sykacak@boku.ac.at	Boku University Vienna, Austria
Stefanie	Tauber	stefanie.tauber@meduniwien.ac	Medical University of Vienna, Austria
Weida	Tong	weida.tong@fda.hhs.gov	Toxicoinformatics FDA, U.S.A.
Thomas	Tuechler	thomas.tuechler@boku.ac.at	Boku University Vienna, Austria
Florian	Van Bömmell	florian.boemmel@charite.de	Charité, Germany
Klemens	Vierlinger	klemens.vierlinger@arcs.ac.at	ARCS, Austria
David	Wild	d.l.wild@warwick.ac.uk	Warwick Systems Biology Centre, U.K.
Ernst	Wit	E.C.Wit@rug.nl	University of Groningen, The Netherlands