Stochastic Centre Workshop in

Statistics for Gene and Protein Expression

May 10 - 12, 2006 at Nya Varvet, Göteborg

Introduction

Identification of factors that influence gene and protein expression are fundamental problems in molecular biology. High throughput techniques such as microarrays, two-dimensional electrophoresis and mass spectrometry pose a number of fascinating and challenging statistical problems in experimental planning and data analysis. The statistical analysis in these areas, often in close cooperation with scientists from biology and medicine, have since the late nineties been highly active research fields.

Practical information

30 minutes are reserved for each talk, followed by 5 minutes for questions. The open pre-workshop seminars will take place at the department of mathematical sciences at Chalmers, lecture hall Euler, on Tuesday May 9. The main workshop will be located at Nya Varvet, by the sea in Gothenburg. For more information visit:

`http://www.math.chalmers.se/Centres/SC/SGPE2006/`

Programme

Tuesday May 9: Open pre-workshop seminars

13:00-13:45 Petter Mostad (docent promotion lecture): Finding the needle in the haystack: Multiple testing in biological experiments

14:00-14:40 Jane Fridlyand: Introduction to the analysis of the array CGH data

14:40-15:10 Coffee

15:10-15:50 Alexander Ploner: Adapting ANOVA for detecting informative peaks in protein mass spectrometry data

16:00-16:40 Mathisca de Gunst: Modelling and analysis of spatio-temporal activity patterns in neuronal networks

Wednesday May 10

09:30-10:00 Coffee

10:00-10:15 Welcome

10:15-10:50 Sylvia Richardson: Bayesian inference in differential expression experiments

10:50-11:10 Coffee

11:10-11:45 Yudi Pawitan: Multidimensional local false discovery rate

11:55-12:30 Natalie Thorne: Issues in the analysis of methylation array data

12:30-14:15 Lunch

14:15-14:50 Claus-Dieter Mayer: Detecting heterogenous variance-covariance structures in gene expression data

14.50-15.15 Coffee

15.15-15.50 Anne-Mette Hein: Aspects of Bayesian gene eXpression (BGX): inference without replicates and accounting for probe affinity effects

16.00-16.35 Ingrid Lönnstedt: Normalization and expression changes in predefined sets of proteins using 2D gel electrophoresis: A study of L-DOPA treated Parkinsonian macaques

17.00-20.00 Poster session

18.00- Food and beverages

Thursday May 11

09.00-09.35 Gordon Smyth: Empirical array quality weights for microarray data

09.45-10.20 Rolf Sundberg: Real-time RT-PCR

10.20-10.50 Coffee

10.50-11.25 Jane Fridlyand: Combining copy number and gene expression data for the analysis of cancer data

11.35-12.10 Eivind Hovig: A sequence oriented comparison of gene expression measurements across different hybridization-based technologies

12.10-13.50 Lunch

13.50-14.25 Søren Bak: Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal inadvertent effects on the metabolome and transcriptome

14.35-15.10 Anders Blomberg: Do you want to have one beer or two? – Proteomics of lager beer yeast strains

15.10-15.40 Coffee

15.40-16.15 Margareta Jernås: Navigating in the fat tissue transcriptome

16.25-17.00 Per Sunnerhagen: Array analysis of post-transcriptional regulation

18.30-19.45 Cultural event

19.45- Workshop dinner

Friday May 12

09.00-09.35 Arnoldo Frigessi: Covariate modulated false discovery rate

09.45-10.20 Claus Ekstrøm: Use of within-array and dye swap replicates for expression analysis in spotted microarrays

10.20-10.50 Coffee

10.50-11.25 Jim Graham: Spot Modeling and Image Registration for Analysis of 2D Electrophoresis Gels

11.35-12.10 Bjarne Ersbøll: Identification of proteins using 2D-electrophoretic gels detecting epilepsy with rats: A comparison of dimension reductive methods

12.10-13.40 Lunch

13.40-14.15 Chris Glasbey: Warping of electrophoresis gels using generalisations of dynamic programming

14.25-15:00 Markus Ringnér: Folding of untranslated regions impact post-transcriptional regulation in yeast

15.00- Coffee

Talk Abstracts

Tuesday May 9: Open pre-workshop seminars

Docent promotion lecture: Finding the needle in the haystack - Multiple testing in biological experiments

Petter Mostad[1]

[1] *Institute of Health Management and Health Economics, University of Oslo, Norway*

In modern medical and biological research, results are often obtained through complex experiments where inference about thousands of variables or hypotheses is required. The statistical methods traditionally used in biological research can reasonably handle a couple or a handful of hypotheses, and can give very wrong results if applied uncritically on data from for example microarray experiments.

In this lecture, I will describe some of the approaches that can be used in such contexts, where inference must be made simultaneously for a long list of hypotheses.The emphasis may be on limiting the possibility of a single falsely rejected null hypothesis (controlling the family-wise error rate) or it may be on controlling the rate of such errors (the false discovery rate). Methods may be based on simply counting the number of hypotheses, or on the dependency between the hypotheses, using for example permutations. Another approach is to estimate the probabilities of different combinations of hypotheses and their alternatives, thus avoiding the hypothesis testing framework.

The ideas will be illustrated with various examples from modern biological research, such as microarray experiments, search for regulatory motifs, and EST expression mining.

Introduction to the analysis of the array CGH data

Jane Fridlyand[1]

[1] *Department of Epidemiology and Biostatistics and Comprehensive Cancer Center, UCSF, USA*

Microarray-based Comparative Genomic Hybridization (Array CGH) is a technique that measures DNA copy number changes, and localizes them on the genome. Such copy number aberrations are common in cancer and in many developmental abnormalities. After outlining the technology, we will discuss statistical methods currently used for their analysis, and future directions.

Adapting ANOVA for detecting informative peaks in protein mass spectrometry data

Alexander Ploner[1]

[1] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden*

Mass spectrometry proteomics, esepcially in its incarnation as SELDI MS-TOF, has shown great potential for detecting clinically relevant biomarkers. The pre-processing of the raw data however is still problematic, especially the distinction between spectral peaks that represent proteins and pure noise. The current crop of algorithms is lacking in both sensitivity and specificity and requires expert user supervision, which makes them impractial for the large amounts of data that SELDI MS-TOF can produce. We have therefore developed a signal detection algorithm based on the simple univariate ANOVA test statistic; by smoothing both the residual error variance and the F-statistic suitably, we achieve both superior sensitivity and specificity and a tractable null distribution. The spectral areas identified by this approach can then be subjected to a traditional peak finding algorithm with a minimum of fuss. The method has been implemented as freely available R package ProSpect.

References:
Tan CS, Ploner A, Quandt A, Lehtio J, Pawitan Y. Finding regions of significance in SELDI measurements for identifying protein biomarkers. *Bioinformatics*, 2006, in print

Modelling and analysis of spatio-temporal activity patterns in neuronal networks

Mathisca de Gunst[1]

[1] *Department of Mathematics, Vrije Universiteit, Amsterdam, The Netherlands*

One of the aims of studying brain tissue or neuronal cells in culture is to obtain information on the connectivity structure of the neuronal cells. To this end electrical activity of the cells is recorded. This yields large and complex data sets and the analysis of these data is generally not straightforward. After a short introduction to neuronal networks, a stochastic model for the firing activity of a population of neurons in culture will be presented. One of the parameters of the model is the connectivity structure of the involved cells. Statistical analysis of real and simulated data based on this model will be discussed.

Wednesday May 10

Bayesian inference in differential expression experiments

Sylvia Richardson[1] and Natalia Bochkina[1]

[1] *Centre for Biostatistics, Imperial College, London, UK*

Differential expression is a key question in many microarray studies. In this talk, we consider it from a Bayesian perspective which allows necessary flexibility for modelling diverse sources of variability usually encountered in microarray data. When building a model for differential expression experiments, one key choice is the structure given to the prior distribution for the parameters of interest, the log fold changes δ_g, where g indexes genes.

Classification of genes as differentially expressed can be seen as a decision problem based on posterior outputs of the model. Alternatively the classification can be directly embedded in the prior structure given to δ_g by using a mixture type formulation. In this presentation, we will build on the model proposed in Lewin et al (2006) and discuss how using a non-informative prior for δ_g and a data-related threshold which takes into account the variability of each gene leads to a new type of classification rule that we call *tail posterior probability.* We will discuss properties of this rule, compare it to other methods for identifying differential expressed genes in a Bayesian framework and propose an estimator for the false discovery rate based on tail posterior probabilities (Bochkina and Richardson, 2006). The alternative framework that uses a mixture prior for δ_g with point mass under the null and parametric models for the alternative will also be briefly reviewed.

References:
Lewin A, Richardson S, Marshall C, Glazier A and Aitman T. (2006) Bayesian Modelling of Differential Gene Expression, *Biometrics*, **62**: 1–9.

Bochkina, N and Richardson, S. (2006). Tail posterior probability for inference in pairwise and multiclass gene expression data. Technical report, Imperial College. Available from http://www.bgx.org.uk

The effects of gene correlation on FDR estimation

Yudi Pawitan[1]

[1] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

Background: The false discovery rate (fdr) is a key tool for statistical assessment of differential expression (DE) in microarray studies. It is, however, well known that overall control of the fdr alone is not sufficient to address the problem of genes with small variance, which suffer from a disproportional high rate of false positives. Graphical tools and modified test statistics have been proposed for dealing with this problem, but there is currently no procedure for controlling the fdr directly.

Methods: We generalize the local fdr called fdr2d - as a function of multiple statistics, combining a common test statistic for assessing differential expression with standard error information.

Results: The fdr2d allows an objective assessment of differential expression as a function of gene variability. Furthermore, the fdr2d has comparable performance to other methods that model the variance explicitly or to the theoretically optimal procedure.

Issues in the analysis of methylation array data

Natalie P. Thorne[1,2], Ashraf E. K. Ibrahim[3], James D. Brenton[4] and Simon Tavaré[1,2]

[1] *Computational Biology Group, Department of Oncology, University of Cambridge, Cambridge, UK*

[2] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK*

[3] *Department of Histopathology, University of Cambridge, UK*

[4] *Department of Oncology, University of Cambridge, Cambridge, UK*

DNA methylation is an epigenetic modification that causes methylation of cytosine bases of CpG's in the mamalian genome. Methylation of CpG's in regulatory elements of a gene can affect transcription and aberrant methylation in such regions has been shown to be associated with disease and in particular with risk of onset of cancer.

DNA methylation profiling studies using microarray technology are becoming increasingly popular. Regardless of the approach, most DNA methylation microarray based methods result in log-ratio data that is characteristically asymmetric. The extent of the skewness in the data is affected by the global levels of methylation in the samples studied and real differences in methylation between samples can be removed through inappropriate use of common normalisation procedures.

Our results are based on replicate experiments using different methylation array methods and for a variety of tissue samples. Our findings show that adequate planning and optimisation are needed to establish spiked controls that are trustworthy in DNA methylation array experiments. Otherwise, normalisation, the ability to compare results between array experiments is problematic.

Detecting heteregenous variance-covariance structures in gene expression data

Claus-Dieter Mayer[1]

[1] *Biomathematics & Statistics Scotland*

Testing for differential expression in gene expression experiments has been one of the most discussed areas within the field of microarray statistics. In the simplest case of comparing gene expression between two groups or experimental conditions t-type tests are commonly used, i.e. the test statistic is given by an appropriately standardized difference of average log-expression in each group. To avoid parametric assumptions permutation methods are often used to calculate p-values or false discovery rates for single genes as well as for the multiple testing of all genes simultaneously. In this talk we will discuss that both (permutation methods and t-type tests) approaches are questionable for genes whose expression distribution changes in more complex way then just a simple shift. We will argue that such complex changes must be expected in biologically interesting situations. Detecting these changes thus is important in two ways: a) it shows whether the use of the traditional methods is valid, b) it can indicate biological information. We will particularly focus on tests to check for changes in the variance-covariance structure of the multivariate gene expression distribution between two or more groups. Tests that that deal with the global multivariate testing problem will be discussed, where dimension reduction by a singular value decomposition (SVD) of the original data matrix allows to use resampling methods in an effective way. We will also indicate how particular genes (pair of genes) with interesting changes in their variance (correlation) can be detected.

Aspects of Bayesian gene eXpression (BGX): inference without replicates and accounting for probe affinity effects

Anne-Mette Hein[1] and Sylvia Richardson[1]

[1] *Centre for Biostatistics, Imperial College, London, UK*

BGX (Hein et al, 2005) is an integrated approach to the analysis of Affymetrix GeneChip arrays. The approach relies on a Bayesian hierarchical model for probe level GeneChip data. Background correction, gene expression level estimation and assessment of differential expression are performed simultaneously. Full posterior distributions of the model parameters can be obtained through MCMC techniques. We explore two aspects of the BGX model: the possibility of performing differential gene expression analysis without replicates and the refinement of the modelling of the non-specific hybridization component to account for probe affinity effects. Differential expression is assessed by comparing the obtained set of posterior probabilities of negative difference in expression P(dg ¡ 0), to that expected under the null hypothesis of no differential expression. The distribution under the null is estimated empirically, by adopting an approach similar to that of Efron (2003). For the second aspect, the refinement of the model is obtained by allowing probe affinity specific distributions of non-specific hybridization. We evaluate the proposed methods on spike-in data sets.

References:
Hein, A. K., Richardson, S., Causton, H. C., Ambler, G. K., and Green, P. J. (2005). BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. *Biostatistics*, 6: 349-373.

Efron, B. 2003. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J Am Statist Assoc*, 99: 96-104.

Hein, A. K. and Richardson, S. (2006). A powerful method for detecting differentially expressed genes from GeneChip arrays with no replicates. Technical report, Imperial College. Available from http://www.bgx.org.uk

Normalization and expression changes in predefined sets of proteins using 2D gel electrophoresis: A study of L-DOPA treated Parkinsonian macaques

Ingrid Lönnstedt[1]

Mathematical Statistics, Uppsala University

The M. J. Fox project aims at investigating changes in protein expression due to acute and chronic L-Dopa treatment in Mptp induced Parkinsonian macaque cells. This first part of the project we base on a study of 2-dimensional difference gel electrophoresis (2D-DIGE).

We have evaluated the performance of different normalization methods for the 2D-DIGE system and also developed a method to assess Differential Expressionin Predefined Protein Sets (DEPPS). The work is joint with Kim Kultima, Uppsala University.

Thursday May 11

Empirical array quality weights for microarray data

Gordon Smyth[1]

[1] *Walter and Eliza Hall Institute of Medical Research, Victoria, Australia*

The accuracy of microarray gene expression data is often questioned, and it is difficult to rigorously verify the accuracy of any high-throughput technology except for a small number of genes which can be exhaustively tested. Meanwhile there is a growing realization that different gene expression platforms may give systematically different results for the same genes, for example because of splice variants.

One approach to evaluating microarray accuracy for all probes simultaneously is to construct a series of RNA samples with a known relationship. Then the precision and sensitivity of the platform can be evaluated by nonlinear regression. This talk will discuss the results of an experiment, using a mixture series of RNA samples, to evaluate and compare four different microarray platforms.

Real-time RT-PCR

Rolf Sundberg[1]

[1]*Department of Mathematics, Stockholm University, Sweden*

A short introduction to real-time RT-PCR type of data will be given. Applications typically involve more samples than genes; there are statistical aspects concerning experimental allocation, modelling of data, and inference. Particular attention will be given to a study of schizophrenia patients versus controls.

Combining copy number and gene expression data for the analysis of cancer data

Jane Fridlyand[1]

[1]*Department of Epidemiology and Biostatistics and Comprehensive Cancer Center*

The development of solid tumors is associated with acquisition of complex genetic alterations, indicating that failures in the mechanisms that maintain the integrity of the genome contribute to tumor evolution. Thus, one expects that the particular types of genomic derangement seen in tumors to reflect underlying failures in maintenance of genetic stability, as well as selection for changes that provide growth advantage. In order to investigate genomic alterations we are using BAC microarray-based comparative genomic hybridization (array CGH). Transcriptional profiles are measured using HGU133A Affymetrix chips. The computational task is to map and characterize the number and types of copy number alterations present in the tumors, and so define copy number phenotypes as well as to associate them with known biological markers and with gene expression data. We define distinct types of genomic events and identify the groups of genes associated with different instabilities. We conclude that various types of genomic instability is associated with the defects in distinct functional groups as determined by Gene Ontology. This result has implications for potential targeted therapies. Additionally, we introduce a graph-theoretic approach that explores relationship between expression, copy number and phenotype in the known pathways.

A sequence oriented comparison of gene expression measurements across different hybridization-based technologies

Eivind Hovig[1]

[1] *Department of Informatics, University of Oslo, Norway*

Gene expression microarrays have made a significant impact in many areas of research. The diversity of platforms and analytical methods has made comparison of data from multiple platforms very challenging. In this study, we describe a framework for cross-platform and cross-laboratory comparisons. We have attempted to include nearly all the available commercial and "in-house" platforms. Probe sequences matched at the exon level across the different microarray platforms produced relatively consistent measurements, in contrast to annotation-based matches. High consistency was seen for highly expressed genes in most platforms, and to a lesser extent for genes with lower expression values as confirmed by QRT-PCR. Measurements obtained using the same technology across laboratories were found to be more concordant than those measured across platforms. We demonstrate that, after stringent pre-processing, (1) commercial arrays were more consistent than "in-house" arrays, and (2) by most measures, one-dye platforms were more consistent than two-dye platforms.

Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal inadvertent effects on the metabolome and transcriptome

Søren Bak[1], Marc Morant[1], Claus T. Ekstrøm[2], Mats Rudemo[2], Carl Erik Olsen[2]

[1] *Plant Biochemistry Laboratory, Department of Plant Biology*

[2] *Department of Natural Sciences, Royal Veterinary and Agricultural University, Copenhagen, Denmark.*

Focused and non-targeted approached were used to assess the impact associated with introduction of new high flux pathways in *Arabidopsis thaliana* by genetic engineering. Transgenic *A. thaliana* plants expressing the entire biosynthetic pathway for the tyrosine derived cyanogenic glucoside dhurrin as accomplished by insertion of three *Sorghum bicolor* cDNAs accumulated up to 4% dry-weight dhurrin with marginal effects on the plant morphology, free amino acid pools, metabolome and transcriptome. In contrast, plant where the incomplete pathway dhurrin pathway was inserted or where biosyntethic pathways were disrupted by mutagenesis accumulated expected as well as unexpected alterations in the metabolome and transcriptome.

Do you want to have one beer or two? - Proteomics of lager beer yeast strains

Anders Blomberg[1]

[1] *Department for Cell and Molecular Biology, Göteborg University, Sweden*

Yeasts are presumably the oldest commercially cultured microorganism and are widely used in the food and beverage industries. There are at least 1000 separate strains of the yeast species *Saccharomyces cerevisiae* currently being commercially used in baking, brewing, distilling and wine-making. The high fermentative capacity of yeasts, together with their ability to withstand the extreme environmental conditions experienced during industrial fermentations, has led to the selection of strains with unique characteristics. Proteome analysis of the three different industrial lager strains revealed the protein content of these strains to be qualitatively rather similar, while they differ substantially to the *S. cerevisiae* laboratory strain. Protein spots in the two-dimensional electrophoresis pattern of the lager strains were subjected to tandem mass spectrometry (LC-MS/MS) based identification. This analysis indicating that proteins in the lager strains that were not found in the 2D pattern of the laboratory strain were most closely related to the corresponding proteins from another yeast species namely *Saccharomyces bayanus*. For many proteins the regulation of these *bayanus*-like proteins and their cerevisiae counterparts varied in a strain dependent manner. The phospoproteome of the lager strains were characterised using the Pro-Q Diamond stain. We found four novel phospho-proteins, Rsp12p, Efb1p, Rsp5p and Leu1p, but no qualitative differences in phosphorylation between the lager strains. In addition, no difference in protein N-terminal acetylation status was observed, generally indicating protein modifications to be of minor importance for the performance of lager strains. Recent results on the differential regulation of the *cerevisiae* and *bayanus* like proteins during various stress conditions will be presented. These differences might influence the robustness of these industrial strains to the industrial process and indicate why these hybrid variants have been selected for commercial use.

References:

Bond, U., and Blomberg, A. (2006) Principles and applications of genomics and proteomics in the analysis of industrial yeast strains Chapter in: Yeast in Food and Beverages, p.175-214 eds. Querol, A., and Fleet, G.H.; Springer Verlag, Berlin, Germany

Navigating in the fat tissue transcriptome

Margareta Jernås[1]

[1] *Department of Metabolism and Cardiovascular Research, RCEM, Göteborg University, Sweden*

Background

Exploring genes and molecular mechanisms involved in the development of complex disorders such as obesity (overweight), is challenging. To increase our knowledge about gene expression in human adipose (fat) tissue we analyzed global gene expression in adipose tissue samples under several different conditions.

Method

Expression profiling using Affymetrix DNA microarray has been performed in our lab. Human adipose tissue has been studied from different aspects such as adipocyte (fat cell) size, adipose tissue distribution, adipose tissue heterogeneity, diet-induced weight loss and hormonal changes.

Results

By combining data from our different projects an overall view of the expression in adipose tissue of a specific gene can be obtained. The power of this approach is here illustrated with the leptin gene. Leptin is a hormone synthesized and secreted primarily by adipocytes and plays a key role in regulating energy intake and energy stores. In our data, leptin was overexpressed in large adipocytes compared to small (3.1-fold). No change was observed between different depots (omental and subcutaneous) of adipose tissue. When adipocytes and stroma-vascular cells were separated, leptin was 4-fold overexpressed in the isolated adipocytes. Leptin expression was downregulated during diet induced weigth loss (2.5-fold). However, there were no changes in leptin expression in adipose tissue from pre- and postmenopausal women.

Conclusion

Our approach allows us to quickly obtain information about the regulation of genes expressed in human adipose tissue. This provides new insights into the physiology and pathophysiology of obesity and associated diseases.

Array analysis of post-transcriptional regulation

Per Sunnerhagen[1]

[1] *Department of Molecular Biology, Göteborg University, Sweden*

Control of gene expression occurs on multiple levels: transcriptional, post-transcriptional, and post-translational. Much emphasis has been put on understanding the mechanisms regulating initiation of transcription on one hand, and regulating protein activity via covalent modifications and protein turnover on the other. More recently, attention has also turned to the second level, control of mRNA translation and stability. In our laboratory, we study both these aspects on the global level using DNA arrays. This poses challenges for data treatment, since the situations where mRNA is collected are clearly non-standard: some of the assumptions underlying the standard models for data normalisation do not hold. I will discuss the issues we want to address relating post-transcriptional control and stress signalling, using yeast as a genetically tractable model.

Friday May 12

Covariate modulated false discovery rate

Egil Ferkingstad, Arnoldo Frigessi[1], Gudmar Thorleifsson, Augustine Kong

[1] *Department of Biostatistics, University of Oslo*
deCode genetics, Reykjavik

Huge amounts of simultaneous comparisons are necessary to extract biological hypothesis from genetic and genomic data. Such comparison tests are dependent, and the dependency structure is unknown, so that the *effective number* of independent tests is unknown. Often, we expect that only a small subset of comparisons will have a positive result: the solution is sparse in the huge parameter space. To discover these solutions, it is necessary to develop new methods that either exploit available a priori knowledge on the structure of sparsity, or merge different data sets, each adding information. Benjamini and Hochberg's false discovery rate (FDR), adapts automatically to sparsity and has been shown to be asymptotically optimal in a certain minimax sense. Efron has developed the theory of local false discovery rate, defined as the probability that the null hypothesis is true given data, casting the testing exercise in an empirical Bayesian setting. We extend this further and introduce the covariate-modulated false discovery rate (cmFDR), useful when there is available a known covariate X_i for each null hypothesis H_{0i} which influences the prior probability that H_{0i} is true. The cmFDR takes advantage of prior information on the probability of each null hypothesis being true based on external additional data, to produce a more precise list of selected genes. This leads to a measure of the posterior significance of each test conditionally on the covariate and the data, possibly leading to greater power. We estimate the cmFDR with help of MCMC and an approximate model on p-values. The new method is applied to the analysis of expression quantitative trait loci (eQTL) data, where gene expression analysis using microarrays is combined with genetic linkage analysis. An eQTL is a genetic variant that influences gene expression. The aim is to test each putative eQTL, and the covariate influencing each test is the estimated heritability of the expression of the corresponding gene. Our method provides a simple way of incorporating this additional information into the data analysis.

Use of within-array and dye swap replicates for expression analysis in spotted microarrays.

Claus T. Ekstrøm[1], Mats Rudemo[1], Marc Morant[2], Søren Bak[2]

[1]*Department of Natural Sciences, KVL.*

[2]*Plant Biochemistry Laboratory, Department of Plant Biology, Center for Molecular Plant Physiology, KVL.*

Dye swap designs and duplicate or triplicate printing are often used for spotted microarrays. The combination of dye swaps and multiple prints makes it possible to partition the variance in both within-array variation (due to multiple printing) and within biological sample variation (due to dye swaps) as well as the normal biological variation. However, the number of biological and technical replicates are often small for microarray experiments so the precision of these variance estimates will be low for a single gene.

We extend the idea from the LIMMA package of having a single common within-array correlation for all genes to accommodate both multiple sources of technical replicates and biological replicates. The extended method either assumes that only the within-array variation is identical for all genes or alternatively that also the within biological sample variation is shared between all genes.

Spot modeling and image registration for analysis of 2D Electrophoresis Gels

Jim Graham[1] and Mike Rogers[1]

[1] *Imaging Science and Biomedical Engineering, The University of Manchester, UK*

2D gel electrophoresis is the most well-established and widely used analytical tool in proteomics. Typically one is seeking to determine changes in the protein complement due to genetic, post-transcriptional or environmental factors. These show up as changes in the size or position of spots in comparative gels. Control samples may be compared against a number of experimental samples, each sample being represented by a number of replicate gels. In addition to meaningful differences, the intensities and positions of spots can vary between gel runs due to a range of artefactual factors, including non-uniformities in the gels or differential heating effects. The functionality of commercially available software for analysing has, in the past, been insufficient to overcome these problems and achieve automatic analysis. The identification of corresponding spots on pairs of gels often requires lengthy, detailed interaction using image analysis packages. Despite recent improvements in image analysis software, and in the reproducibility of gels themselves, gel analysis still requires considerable user input to reach a satisfactory comparison. Achieving a consistent analysis across a set of gels is extremely difficult. We have investigated two aspects of the gel analysis problem: spot modeling and gel registration. Spot modeling involves characterisation of spots for identification and quantitative description. A number of spot models are used. Parametric methods, using, for example, gaussian or diffusion models, make strong assumptions about spot appearance and are often insufficiently flexible to adequately represent all spots that may be present in a gel. Nonparametric methods make no assumptions about spot appearance and consequently impose few constraints on spot detection, allowing more flexibility, but reducing robustness when image data is complex. We have investigated an approach using a statistical model of shape, based on the statistics of an annotated training set. The model allows new spot shapes, belonging to the same statistical distribution as the training set, to be generated. To represent a spot surface patch we use the statistically derived shape convolved with a Gaussian kernel, simulating the diffusion process in spot formation. The statistical model of spot parameters shows both greater accuracy of fit and higher specificity (distinction of multiple spots from single spots) than parameterisations based solely on Gaussian and diffusion models. In gel registration we have adopted a point-matching approach. We conducted an extensive investigation of point-matching methods, evaluating their performance in the presence of large image distortions and errors in spot detection, both false positive and false negative. We adopted a method based on iterated closest point matching, using robust calculation of point correspondences and a distance metric derived from image structure. Our evaluation shows high accuracy and robustness even in the presence of large image distortions and proportions of "outlier" spot detections. This opens the possibility of fully automatic registration of gels across sets, rather than interactive alignment of pairs.

Identification of proteins using 2D-electrophoretic gels detecting epilepsy with rats: A comparison of dimension reductive methods

Line H. Clemmensen[1] and Bjarne K. Ersbøll[1]

[1] *Informatics and Mathematical Modelling, Technical University of Denmark*

Dye swap designs and duplicate or triplicate printing are often used for spotted microarrays. The combination of dye swaps and multiple prints makes it possible to partition the variance in both within-array variation (due to multiple printing) and within biological sample variation (due to dye swaps) as well as the normal biological variation. However, the number of biological and technical replicates are often small for microarray experiments so the precision of these variance estimates will be low for a single gene.

We extend the idea from the LIMMA package of having a single common within-array correlation for all genes to accommodate both multiple sources of technical replicates and biological replicates. The extended method either assumes that only the within-array variation is identical for all genes or alternatively that also the within biological sample variation is shared between all genes.

References:
Zou H. & Hastie T. (2005), Regularization and variable selection via the elastic net, *J. R. Statist. Soc. B* **67** (Part 2), 301-320.

Warping of electrophoresis gels using generalisations of dynamic programming

Chris Glasbey[1]

[1]*Biomathematics and Statistics Scotland*

Dynamic programming (DP) is a fast, elegant method for finding the global solution to a class of optimisation problems. For example, it can be used to align pairs of tracks in 1-D electrophoresis gels such as pulsed-field gel electrophoresis (PFGE), which is used to genotype bacterial samples such as E. coli O157 strains by DNA fingerprinting. However, it is not possible to use DP to align many PFGE tracks or to align pairs of 2-D polyacrylamide gels (SDS-PAGE), which are used, for example, to distinguish between strains of malarial parasite.

We consider three generalisations of DP for alignment or warping of 1-D and 2-D gels. The first approach is a greedy algorithm first proposed by Leung et al (2004), termed iterated dynamic programming (IDP), where DP is used to recursively solve each of a sequence of lower-dimensional problems in turn, to find a local optimum. A second algorithm replaces DP by a more computationally intensive Forward-Backwards Gibbs Sampler (Scott, 2002), and uses a simulated annealing cooling schedule to guarantee the optimal solution. The final approach is an empirical, stochastic optimiser, which is implemented by adding noise to IDP. Methods are illustrated using PFGE and SDS-PAGE data.

References:
Leung, C., Appleton, B. and Sun, C. (2004). Fast stereo matching by Iterated Dynamic Programming and quadtree subregioning. *British Machine Vision Conference* (Eds. A Hoppe, S Barman and T Ellis) 1, 97-106.

Scott, S.L. (2002). Bayesian methods for Hidden Markov Models: recursive computing in the 21st century. *Journal of the American Statistical Association*, 97, 337-351.

Folding of untranslated regions impact post-transcriptional regulation in yeast

Markus Ringnér[1]

[1]*Department of Theoretical Physics, Lund Univeristy, Sweden*

Using high-throughput technologies, abundances and other features of genes and proteins have been measured on a genome-wide scale in Saccharomyces cerevisiae. In contrast, secondary structure in untranslated regions (UTRs) of mRNA has only been investigated for a limited number of genes. Here, we present a study of genome-wide regulatory effects of mRNA 5' UTR folding free energies. We performed computations of secondary structures in 5' UTRs and their folding free energies for all verified genes in S. cerevisiae. We found significant correlations between folding free energies of 5' UTRs and various transcript features measured in genome-wide studies of yeast. In particular, mRNAs with weakly folded 5' UTRs have higher translation rates, higher abundances of the corresponding proteins, longer half-lives, and higher numbers of transcripts, and are upregulated after heat shock. Furthermore, 5' UTRs have significantly higher folding free energies than other genomic regions and randomized sequences. We also found a positive correlation between transcript half-life and ribosome occupancy that is more pronounced for short-lived transcripts, which supports a picture of competition between translation and degradation. Among the genes with strongly folded 5' UTRs, there is a huge overrepresentation of uncharacterized open reading frames. Based on our analysis, we conclude that (i) there is a widespread bias for 5' UTRs to be weakly folded, (ii) folding free energies of 5' UTRs are correlated with mRNA translation and turnover on a genomic scale, and (iii) transcripts with strongly folded 5' UTRs are often rare and hard to find experimentally.

Poster Abstracts

P1. Fully Bayesian mixture model for gene expression

Alex Lewin[1]

[1] *Imperial College, London, UK*

Mixture models are commonly used to classify genes in differential expression experiments. Often the proportion of differentially expressed genes is fixed or empirically estimated when fitting these models. Here we show that this proportion, and the mixture parameters, can be e stimated in a fully Bayesian way. We use a log-linear model at the data level, and a mixture on the differential expression parameters of a point mass at zero, corresponding to the null hypothesis, and two separate distributions for over and under-expressed genes. The model is i mplemented in C++ using reversible-jump MCMC, in order to correctly estimate the mixture of a point mass and a continuous distribution.

We classify the genes as null, over or under-expressed according to their posterior probabilities of being in the different mixture componen ts. This enables us to obtain a very straightforward estimator for the false discovery rate. Simulation studies show that the model can esti mate both the proportion of true nulls and the false discovery rate well. We explore the performance of the model in a variety of situations and investigate the robustness of the estimation to the choice of prior on the differential expression parameters

P2. The Hotelling T-test for testing pathways in microarray gene expression data

Alexander Ploner[1]

[1] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden*

In the context of functional genomics, there is growing interest in integrating existing biological knowledge about gene interactions or gene functions into the analysis of microarray expression data. Many recent approaches in this area test in some way or other for enrichment of functional categories among differentially expressed genes, but a few researchers have adapted traditional multivariate techniques like the Hotelling T-test to study regualtion of whole gene pathways.

In our poster, we describe how the number of samples and genes influences the

estimation the gene-gene covariance matrix and the power of the Hotelling T-test. We demonstrate that a rank-reduced version of the covariance matrix can improve power, and present an Omnibus test that uses all possible rank reductions and still manages to run in manageable time.

P3. Genome-wide analysis of differential mRNA splicing based on Affymetrix All Exon Arrays

Anja von Heydebreck[1]

[1] *Department of Bio- and Chemoinformatics, Merck KGaA, Germany*

Affymetrix All Exon Arrays aim at measuring mRNA abundances of all human exons, which makes them a promising tool to study alternative splicing in human tissue and cell culture samples. Using publicly available data, we will discuss the preprocessing of Exon Array data as well as statistical methods that may be applied to identify tissue-specific splice variants.

P4. The influence of missing value imputation on detection of differentially expressed genes

Ida Scheel, Magne Aldrin, Ingrid Glad, Ragnhild Sørum, Heidi Lyng and Arnoldo Frigessi

Department of Mathematics and Department of Biostatistics, University of Oslo
Norwegian Computing Centre
The Norwegian Radium Hospital

Missing values are problematic for the analysis of microarray data. Imputation methods have been compared in terms of the similarity between imputed and true values in simulation experiments, while their influence on the final results has not been studied. Also, the focus has been on missing at random, while entries are missing also not at random. We investigate the influence of imputation on the detection of differentially expressed genes from cDNA microarray data. We apply ANOVA models and the popular SAM FDR-based testing procedures and look to the differentially expressed genes that are lost because of imputation. This way of looking to the success of imputation provides useful information that the traditional root mean squared error cannot capture. We also show that the type of missingness matters: imputing 5% missing not at random has the same effect as imputing 10-30% missing at random. Finally, we propose a new method for imputation (LinImp), fitting a simple linear model for each channel separately, and compare it with the widely used KNNimpute method. For 10% missing at random, KNNimpute leads to twice as many lost differentially expressed genes as LinImp

P5. Bayesian networks for biological networks exploration using publicly available expression data

Darima Lamazhapova[1]

[1]*Lundberg Laboratory for Cancer Research, Department of Pathology, Göteborg University, Sahlgrenska University Hospital, Sweden*

Bayesian network learning method was implemented for structure recovery for small datasets. The method relies on estimation of statistical confidence in the some "features" of the network under consideration using non-parametric bootstrap method. It was evaluated on simulated data from ALARM network. Further, gene expression datasets were retrieved from publicly available databases and the method was applied for structure learning of the following biological networks: SRF transcription factor related network, smooth muscle cells regulation network, gene module network.

P6. Estimation of finite mixtures for SNP genotyping

Hedvig Norlén[1]

[1]*Department of Mathematics, Stockholm University, Sweden*

Single nucleotide polymorphisms (SNPs) are bi-allelic single base changes with a frequency of approximately one in every 300 bp. In some cas es SNPs have been demonstrated to be associated with specific disease states, such as Sickle cell anemia. In spite of this, in the vast majority of cases it is unlikely that a SNP can be causally linked to a certain trait. Instead, certain combinations of variants among the millions of SNPs in the genome are likely to give rise to traits such as high susceptibility to cancer or adverse reaction to drugs. To ascertain the relationship between genetic variations and such phenotypic differences a large number of SNPs must be analyzed in a large number of individuals. Consequently, accurate and sensitive high-throughput scoring techniques will play an important role in mapping disease markers and in routine diagnostics. We have developed a statistical technique for mixture identification and classification, for use with a novel microarray technique for SNP genotyping, based on allele-specific primer extension. The aim is an accurate, rapid and largely automated genotyping methology.

P7. Maximum likelihood classifiers in microarray studies

Jens Ledet Jensen[1]

[1]*Department of Theoretical Statistics, University of Aarhus*

The use of microarray data for classification purposes is widespread. In such applications one faces the "curse of dimensionality" in a very direct way: of the many variables (genes) measured only a small number actually show differential expression between the two groups of interest. Thus, most variables simply add noise to a classifier and variable selection becomes an important issue. In this presentation a number of simple illustrations of the problem and its possible solution through thresholding are given. In particular the case of unequel sample sizes is discussed and the method of shrunken centroids is put into the general framework of thresholding.

P8. Transcriptomic and metabolomic characterization of *rnt1-1*, an *Arabidopsis thaliana* knockout in a metabolic branch point between primary and secondary metabolism

Marc Morant[1], Claus Ekstrøm[2], Carl Erik Olsen[2], Birger Lindberg Møller[1], Mats Rudemo[2] and Søren Bak[1]

[1] *Plant Biochemistry Laboratory, Department of Plant Biology, Center for Molecular Plant Physiology, Royal Veterinary and Agricultural University, 40 Thorvaldsensvej, DK-1871 Frederiksberg C, Copenhagen, Denmark*

[2] *Department of Natural Sciences, Royal Veterinary and Agricultural University, Denmark*

In the *Arabidopsis thaliana* knockout mutant *rnt1-1* a key enzyme situated at a metabolic branch point between primary and secondary metabolism is perturbed. *rnt1-1* seedlings exhibit a characteristic auxin overproduction phenotype as characterized by elongated hypocotyls, epinastic cotyledons and proliferation of hairy secondary roots. RNT1-1 encodes CYP83B1, a cytochrome P450 known to be involved in biosynthesis of glucosinolates, a class of secondary metabolites present in *A. thaliana*. However, the visual phenotype shown that in *rnt1-1* homeostasis of the phytohormone auxin is also perturbed.

We describe the impact on the transcriptome and metabolome of 10 days old *rnt1-1* seedlings; a developmental stage where the plant visual phenotype is still comparable. Three biological replicates were analyzed using Agilent commercial 22K global *A. thaliana* 60mer oligonucleotide arrays and a custome made 50mer spotted oligonucleotide array designed to detect 455 selected genes in primarily secondary metabolism. To select differentially regulated genes in *rnt1-1*, Imagene and ImageSight (Biodiscovery) were used for image analysis and data normalization followed by SAM analysis (Significant Analysis for Microarray). In parallel, hot methanol extracts were analyzed by untargeted LC-MS (Liquid Chromatography-Mass Spectrometry) metabolite profiling and the MetAlign software was used to select significant alterations in the *rnt 1-1* metabolome. Our experimental design allows for untargeted as well as targeted analysis for alterations in the metabolome and transcriptome. Future experiments will correlate the changes in the metabolome and transcriptome by multivariate data analysis, and selected genes will be heterologously expressed and biochemically characterized.

P9. Using large-scale metabolite and gene-expression analyses in rainbow trout to identify responses to pharmaceuticals

Lina Gunnarsson[1], Linda M Samuelsson[1], Lars Förlin[2], Göran Karlsson[3], Margaretha Adolfsson-Erici[4], Erik Kristiansson[5], Olle Nerman[5] and Joakim Larsson[1]

[1] *Department of Neurosciences and Physiology, the Sahlgrenska Aacademy at Göteborg University, Göteborg, Sweden.*

[2] *Department of Zoology/Zoophysiology, Göteborg University, Göteborg, Sweden.*

[3] *Swedish NMR Centre at Göteborg University, Sweden.*

[4] *Department of Applied Environmental Science (ITM), Stockholm University, Stockholm, Sweden.*

[5] *Department of Mathematical Statistic, Chalmers University of Technology, Göteborg, Sweden*

The recent advances in the characterization of genomes and development of high-throughput screening methods and bioinformatics have opened up previously unexplored and powerful ways to approach key environmental issues. By analysing hundreds or thousands of potential responses simultaneously, the possibilities to discovering unexpected effects are greatly increased. This approach may become important for the identification of specific and sensitive markers of exposure and adverse effects, understanding toxic mechanisms, directing testing to certain outcomes, identifying sensitive species and environmental monitoring. As a proof of concept, we have applied exploratory molecular analyses on a relatively well characterized model, the estrogen-exposed rainbow trout. Fish were exposed to 0, 0.87 or 10 ng/L of ethinylestradiol (measured concentrations) for two weeks in a flow-through system. Blood plasma was analyzed with 1H-NMR for metabolic effects and microarrays were used to characterize gene expression in liver. Altered levels of metabolites, such as alanine, phospholipids, cholesterol and vitellogenin residues were found in plasma of fish exposed to 10ng/L. In accordance, the presence of vitellogenin in plasma was demonstrated in the high-dose group by ELISA and Western blot. Using a salmonid microarray, a large number of regulated genes were identified in the liver, including vitellogenin, vitelline envelope proteins, cathepsin D and fatty-acid binding proteins. Some transcriptional changes were also observed in the low-dose group. Many responses recognized by these exploratory methods could be put in context with previous knowledge on the effects of estrogens in fish and other vertebrates. This adds confidence to the approach of using NMR-metabolomics and microarrays to identify environmental effects of pharmaceuticals and other contaminants.

P10. Comparative genomics study of upstream open reading frames

Marija Cvijovic[1], Per Sunnerhagen[2], Olle Nerman[3], Graham Kemp[4]

[1] *Max Planck Institute for Molecular Genetics, Berlin, Germany*

[2] *Department of Molecular Biology, Göteborg University, Sweden*

[3] *Department of Mathematical Statistics, Chalmers University of Technology, Sweden*

[4] *Computer Science and Engineering, Chalmers University of Technology, Sweden*

Identification of elements responsible for post-transcriptional control and their functional represents a relatively unexplored area of molecular biology. This study shows that using comparative genomics, it will likely be possible to predict which upstream ORFs are functional. Cross-species analysis shows that important parameters for conservation of uORFs can be extracted and used in identification of genes that are regulated on the translational level.

P11. Statistical testing within the gene ontology hierarchy

Clara-Cecilie Günther[1], Mette Langaas[1], Stian Lydersen[2], Vidar Beisvåg[2,3], Frode K. R. Jünge[2,3], Hallgeir Bergum[2,3], Astrid Lægreid[2,3]

[1] *Department of Mathematical Sciences, The Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.*

[2] *Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, NO-7489 Trondheim, Norway.*

[3] *Norwegian Microarray Consortium (NMC).*

The overall aim of Systems Biology is to come to an understanding of complete biological systems.Different sources of data may enter into the modelling of the systems, e.g. microarray experiments for measuring gene expression. A popular aim of gene expression experiments is to arrive at one (or several) set(s) of reporters (genes, probe sets, ESTs) that are found to be differentially expressed between situations (e.g. treatment A vs. treatment B). To aid in the interpretive challenge of summarizing the findings present in the obtained lists of reporters, a strategy called gene-class testing (GCT) has been proposed. Gene classes may be based on Gene Ontology (GO) categories. eGOn (explore Gene Ontology), http://www.genetools.no, is a GO-tool where lists of reporters can be submitted trough a web interface to a annotation database, and automatically translated into GOterms annotated to these reporters. In addition to powerful graphical displays eGOn offers statistical hypothesis testing to assess the level of similarity between two reporter lists. Let us consider two reporter lists; list A and list B. At the given GO-node G, we are interested in testing whether the probability of belonging to GO-node G is different for reporter list A and reporter list B. I.e. for each reporter on list A, there is a probability P(G—A) of belonging to GO-node G, and for each reporter on list B, there is a probability P(G—B) of belonging to GO-node G. Under the null hypothesis these two probabilities are equal. At each gene class (e.g. GO-node) we distinguish between three situation, due to possible dependencies between the lists A and B:

i) One Reporter List is a Subset of the Other List: one of the two reporter-lists compared is the list containg all reporters present at that GO-node in the full experiment. (e.g. all reporters assayed on the chip in microarray experiment).

ii) Mutually Exclusive Reporter Lists: two reporter lists, A and B, are compared, and there are no reporters that are on both lists, e.g. A is a list of reporters associated with up-regulation and B is a list of reporters associated with down-regulation.

iii) Intersecting Reporter Lists: two reporter lists, A and B, are compared, and there exist reporters that are on both lists, e.g. A is a list of reporters associated with treatment A and B is a list of reporters associated with treatment B.

For most GO-tools only the situation (i) above is handeled. For situations (i) and (ii) several approaches can be used, e.g. Fisher's exact test, Pearson's χ^2 statistic, the mid-p test and unconditional tests. The focus of this presentation is on studying and developing new tests for handeling the situation (iii) above. The tests are compared and evaluated in a simulation study, and applied to reporter lists from gene expression studies.

P12. A probabilistic treatment of missing spots in 2D gel experiments

Morten Krogh[1], Celine Fernandez, Maria Teilum, Sofia Bengtsson and Peter James

[1] *Computational Biology & Biological Physics, Lund University, Sweden*

2D gels are widely used to measure the abundances of hundreds of proteins simultaneously. Usually, protein abundances of two or more biological groups are compared using biological and technical replicates. Spots are detected, spot volumes are quantified, and spots are matched across gels. There are almost always many missing values in the resulting data set. The missing values arise either because the protein has very low abundance or because of experimental errors including bad gels and faulty spot detection and matching. In this study, we first show that the probability for a spot to be missing is reasonably well modeled by a logistic regression function of the logarithm of the volume. Then we present an algorithm that takes a set of gels of technical or biological replicates as input and estimates the true protein abundances from the number of missing spots and measured volumes of the present spots using a maximum likelihood approach. Statistical significance is calculated using bootstrap sampling. The algorithm is compared to two standard approaches, and is shown to perform well.

P13. Analysing 1.5-channel microarray data

Andy G Lynch[1,2], David E Neal[3], Glynn T Burtt[3] and Natalie P Thorne[1,4]

[1] *Computational Biology Group, Department of Oncology, University of Cambridge, Cambridge, UK*

[2] *Centre for Applied Medical Statistics, Department of Public Health and Primary Care, Cambridge, UK*

[3] *Department of Oncology, University of Cambridge, Cambridge, UK*

[4] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK*

There are a number of mechanisms that can lead to one channel of a two-channel microarray experiment being partially or completely corrupted. In particular, the Cyanine-5 (red) dye is susceptible to ozone damage that can leave one with only the green channel if experiments were conducted in a high ozone environment.

If only some arrays are affected by damage to one channel, then there is a dilemma with regard to the analysis. Off-the-shelf solutions lend themselves

to either conducting a single channel analysis (i.e. discarding the affected channel even from unaffected arrays) or a two-channel analysis on unaffected arrays (i.e. discarding the unaffected channel from affected arrays). Both of these options are clearly wasteful including the popular third option of ditching the entire dataset and starting again.

P14. BioHMM: a heterogeneous HMM for segmenting array CGH data

John C. Marioni[1,2], Natalie P. Thorne[1,2], Simon Tavaré[1,2]

[1] *Computational Biology Group, Department of Oncology, University of Cambridge, Cambridge, UK*

[2] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK*

We have developed a new method (BioHMM) for segmenting array comparative genomic hybridisation data into states with the same underlying copy number. By utilizing a heterogeneous hidden Markov model, BioHMM incorporates relevant biological factors (e.g. the distance between adjacent clones) in the segmentation process. BioHMM is part of the R library snapCGH which is available from Bioconductor.

We outline the methodology and ideas behind the approach. The algorithm is illustrated using a variety of datasets from different platforms. Some examples are given comparing its performance to other segmentation schemes. Finally we describe the possible extensions to BioHMM such as incorporating clone quality weights.

P15. Low level analysis of Illumina BeadArray data

Mark J. Dunning[1], Natalie P. Thorne[1,2], Mike L. Smith1, Isabelle Camilier[3] and Simon Tavaré[1,2]

[1] *Computational Biology Group, Department of Oncology, University of Cambridge, Cambridge, UK*

[2] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK*

[3] *Ecole Polytechnique, Paris, France*

Illumina BeadArrays provide a novel microarray technology using randomly assembled arrays of beads. Each bead on the array carries copies of a single gene-specific probe with, on average, about 30 replicates of each bead type on an array. Such a large degree of replication on each array and the parallel way in which arrays can be combined make BeadArrays highly suited in high-throughput experiments such as for genome-wide population studies.

We have developed an R library (*beadarray*) which allows BeadArray data to be read and analysed in a more flexible manner than existing software. The library is available on the Bioconductor developer site. We present our results investigating the low level analysis of such data including image analysis, quality assessment, outlier detection, background correction, and normalisation.

P16. P-value estimates in microarray-based regression problems

Patrik Edén[1]

[1] *Department of Theoretical Physics, Lund University, Sweden*

To look for combinations of gene expressions that correlate to the target value under study, one may construct predictors taking gene expression as input, and see if they match the targets better than expected by random. For classification problems, standard techniques are available, e.g. Fisher's exact test on odds ratios. We present an approach to estimate the P-value in regression problems, where the target is a continuous variable rather than a class.

P17. A simple mathematical model of adaptation to high osmolarity in yeast

Peter Gennemark[1], Bodil Nordlander[2], Stefan Hohmann[2] and Dag Wedelin[1]

[1] *Department of Computer Science and Engineering, Chalmers University of Technology, Sweden*

[2] *Department of Cell and Molecular Biology/Microbiology, Göteborg University, Sweden.*

We present a simple ordinary differential equation (ODE) model of the adaptive response to an osmotic shock in the yeast *Saccharomyces cerevisiae.* The model consists of two main components. First, a biophysical model describing how the cell volume and the turgor pressure are affected by varying extra-cellular osmolarity. The second component describes how the cell controls the biophysical system in order to keep turgor pressure, or equivalently volume, constant. This is done by adjusting the glycerol production and the glycerol outflow from the cell. The complete model consists of 4 ODEs, 3 algebraic equations and 10 parameters. The parameters are constrained from various literature sources and estimated from new and previously published absolute time series data on intra-cellular and total glycerol.

The qualitative behaviour of the model has been successfully tested on data from other genetically modified strains as well as data for different input signals. Compared to a previous detailed model of osmoregulation, the main strength of our model is its lower complexity, contributing to a better understanding of osmoregulation by focusing on relationships which are obscured in the more detailed model. Besides, the low complexity makes it possible to obtain more reliable parameter estimates.

P18. Detecting gene expression patterns related to the development of osteoporosis

Petter Mostad[1]
[1]*Institute of Health Management and Health Economics, University of Oslo, Norway*

Osteoporosis is a condition characterized by decreased bone mineral density (BMD),leading to fragile bones and increased risk of fractures. It is quitecommon especially for older women, and large resources are invested in prevention, treatment, and research.

Loss of bone mineral density is a natural part of aging, but there are large individual differences in the speed of this process, and also differences in the maximum BMD reached by adults before deterioration starts.These differences are known to be influenced by genetic factors, although environmental factors are also important. In our study, we focus on gene expression. In an experiment at Rikshospitalet in Oslo, more than 100 women were examined for BMD and other clinical parameters, and for a majority of the women,expression profiling was done on bone biopsies.

An important challenge when analyzing this data is to find not only genes whose expression levels are correlated with low BMD, but also to find those correlated with rapid loss of BMD. Although the first type is important both for diagnosis and for understanding BMD loss, finding genes of the second type would be even more important and interesting for understanding the disease.

As we only have measurements at one timepoint for each woman, we try to use clinical parameters like age, time since menopause and other clinical data to estimate the rate of BMD loss. In this poster, we present some preliminary findings from the data, but focus on the ideas for further analysis.

P19. A test for partial differential expression

Wessel N. van Wieringen[1], Mark A. van de Wiel[1] and Aad W. van der Vaart[1]

[1]*Department of Mathematics, Vrije Universiteit De Boelelaan, Amsterdam, The Netherlands*

In cancer research comparative microarray experiments are among others carried out to identify genes that are differentially expressed between normal and cancer tissue. As cancer of a particular tissue type is often a collection of different diseases, each with its own genetic mechanism, a gene may well be expressed at the same level in the normal and cancer tissues, except it is expressed at a different level for a proportion of the cancer tissues. A well known example is found in breast cancer, where the ERBB2 gene is over-expressed (with respect to normal tissue) in only a proportion of the breast carcinomas. We developed a two-sample test that, as opposed to commonly used tests, is designed to detect shifts that occur in part of one sample only (called partial shifts). Such a test is of particular interest in gene expression studies involving cancer tissue, and may identify the ERBB2 gene where other tests may not. In the construction of the test we modelled the outcome of a two-sample comparative microarray experiment with a non-parametric mixture. The metric of partial differential expression is the mixing component of this mixture. Hence, the test statistic is related to the mixing component, and inspired on the theory of minimum distance estimation. The null-distribution of the test statistic is obtained by permutation re-sampling. The thus proposed permutation test is shown to be asymptotically distribution-free and consistent. An extensive simulation study, covering a wide range of situations, shows that the proposed test is more powerful than the two-sample t-test and the Wilcoxon rank sum test for partial shifts, while it is competitive for whole-sample shifts. Application of the proposed test to a real-life dataset consisting of the expression profiles of normal and ovarian cancer tissue identifies genes that are clearly bimodal, indicating a partial shift. These genes were not identified by the two-sample t-test and the Wilcoxon rank sum test.

P20. Arsenic and the sulphur metabolism

Thorsen M[1], Kristiansson E[3], Lagniel G[2], Nerman O[3],Labarre J[2] and Tamas MJ[1]

[1]*Department of Cell and Molecular Biology/Microbiology, Göteborg University, Sweden*

[2]*Service de Biochimie et de Génétique Moléculaire, DBJC, CEA/Saclay, France*

[3]*Department of Mathematical Statistics, Chalmers University of Technology, Sweden*

Arsenic is widely distributed in nature and all organisms possess regulatory mechanisms to evade toxicity. Arsenic is a well-established human carcinogen but is also used in several medical treatments. The cellular response of S. cerevisiae to arsenic stress is investigated by transcriptome, proteome and metabolome analysis. Arsenite stress give rise to a general up regulation of oxidative stress defence genes and many sulphur related genes, including all the genes encoding functions in the glutathione biosynthesis. We show that under arsenite exposure, gene regulation of the components of the sulphur metabolism are controlled by Yap1, a transcription factor of the oxidative stress response, in addition to Met4, the classical transcription factor of the sulphur metabolism. Glutathione is an essential metabolite in yeast (and in all higher eukaryotes). Induction of GSH1 has previously been reported under arsenite stress and this has been interpreted as a sign of depletion of the cellular glutathione pool. Therefore, depletion of glutathione has been proposed to be the main toxic effect of arsenite. Conversely, by measuring the metabolites in the glutathione metabolism pathway, we can show that the glutathione pool is dramatically increased upon arsenite exposure. Our data shows that arsenite exposed cells re-priorities the sulphur fluxes of the cell; under normal conditions about 90% of assimilated sulphur is channelled into proteins whereas under arsenite stress about 50% of the assimilated sulphur goes into glutathione biosynthesis.

P21. Navigating in the fat tissue transcriptome

Margareta Jernås[1]

[1]*Department of Metabolism and Cardiovascular Research, RCEM, Göteborg University, Sweden*

Background

Exploring genes and molecular mechanisms involved in the development of complex disorders such as obesity (overweight), is challenging. To increase our knowledge about gene expression in human adipose (fat) tissue we analyzed global gene expression in adipose tissue samples under several different conditions.

Method

Expression profiling using Affymetrix DNA microarray has been performed in our lab. Human adipose tissue has been studied from different aspects such as adipocyte (fat cell) size, adipose tissue distribution, adipose tissue heterogeneity, diet-induced weight loss and hormonal changes.

Results

By combining data from our different projects an overall view of the expression in adipose tissue of a specific gene can be obtained. The power of this approach is here illustrated with the leptin gene. Leptin is a hormone synthesized and secreted primarily by adipocytes and plays a key role in regulating energy intake and energy stores. In our data, leptin was overexpressed in large adipocytes compared to small (3.1-fold). No change was observed between different depots (omental and subcutaneous) of adipose tissue. When adipocytes and stroma-vascular cells were separated, leptin was 4-fold overexpressed in the isolated adipocytes. Leptin expression was downregulated during diet induced weigth loss (2.5-fold). However, there were no changes in leptin expression in adipose tissue from pre- and postmenopausal women.

Conclusion

Our approach allows us to quickly obtain information about the regulation of genes expressed in human adipose tissue. This provides new insights into the physiology and pathophysiology of obesity and associated diseases.

P22. Weighted Analysis of Microarray Experiments

Erik Kristiansson[1], Anders Sjögren[1], Mats Rudemo[1], Olle Nerman[1]

[1] *Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden*

In microarray experiments quality often varies, for example between samples and between arrays. The need for quality control is therefore strong. A statistical model and a corresponding analysis method is suggested for experiments with pairing, including designs with individuals observed before and after treatment and many experiments with two-colour spotted arrays. The model is of mixed type with some parameters estimated by an empirical Bayes method. Differences in quality are modelled by individual variances and correlations between repetitions. The method is applied to several real datasets, both of Affymetrix and two-colour cDNA type. In all cases, the patients or arrays had different estimated variances, leading to distinctly unequal weights in the analysis. For simulated data the improvement relative to previously published methods without weighting is shown to be substantial.

References:
Kristiansson, E., Sjögren, A., Rudemo, M., Nerman, O. (2005). Weighted Analysis of Paired Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* Vol. 4: No.1, Article 30.

Kristiansson, E., Sjögren, A., Rudemo, M., Nerman, O. (2006). Quality Optimised Analysis of General Paired Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* Vol. 5: No. 1, Article 10.

P23. The effect of between array correlations - can you trust your p-values?

Anders Sjögren[1], Erik Kristiansson[1], Mats Rudemo[1], Olle Nerman[1]

[1] *Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden*

In microarray analysis, measurements from different arrays are often assumed to have independent noise with equal variance. We examine one group in the well known dataset of Golub et al (1999) and conclude that these assumptions do not hold. Correlations exist between arrays and some arrays have higher variability than others.

To examine the effect of the invalid assumptions, we try to find differentially expressed genes between random subgroups from the same condition, i.e. where we know there should be no differentially expressed genes. The distributions of the resulting p-values are far from what would be expected. The effect would be that the estimates of e.g. the false discovery rate and the number of regulated genes would be highly dependent on unknown random factors. For example, one might get a high estimate of the number of regulated genes when there are none. The Weighted Analysis of Microarray Experiments model (WAME, Kristiansson et al. (2005, 2006)) includes a covariance structure and aims at handling this situation with correlations and different variances.

The power of the tests are examined by adding a known signal to the randomly selected subgroups, thus avoiding simulation of the noise according to some selected model. The result is that WAME performs favourably compared to the alternatives studied. However, WAME gives too conservative p-values in cases where a considerable proportion of genes are regulated.

Conclusion: All of the examined methods give biased p-values, and it seems questionable to trust microarray p-values at the current state of the art.

References:
Golub et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

Kristiansson, E., Sjögren, A., Rudemo, M., Nerman, O. (2005). Weighted Analysis of Paired Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* Vol. 4: No.1, Article 30.

Kristiansson, E., Sjögren, A., Rudemo, M., Nerman, O. (2006). Quality Optimised Analysis of General Paired Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* Vol. 5: No. 1, Article 10.

P24. Normalisation and analysis of a post-transcriptional mRNA assay

Janeli Sarv[1], Swarna Swaminathan[2], Marija Cvijovic[3], Per Sunnerhagen[4], Olle Nerman[1]

[1] *Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden*
[2] *Department of Genetics, Cell Biology, and Development, University of Minnesota at Minneapolis, Minneapolis, USA*
[3] *Max Planck Institute for Molecular Genetics, Berlin, Germany*
[4] *Department of Molecular Biology, Göteborg University, Sweden*

The dataset investigated contains Yeast GeneFilter generated data,repeatedly measured with different preciseness. Some hybridizations were recorded in more than one exposure with varying time length, and in the case of longer exposure times one gets saturations of high signal values. In original analysis all the exposures were treated as independent experiments, even if they represent exactly the same array. Thus, the dependence of exposures from the same array is not taken account of, and variability of the signal intensities between arrays is underestimated. Moreover, a drawback from the software used in the signal extraction is that raw signal data are not available. Here several different modifications of the data to handle those difficulties are suggested.

One of the main goals is to improve the quality of the data and use the corrected data to investigate the set of mRNAs that have upstream open reading frames (uORFs). The expression of those mRNAs might be controlled by ribosome interactions with uORFs and abundance of other regulatory complexes. The array experiments are specially designed to study changes in translation rates under stress response, that might induce such ineractions.

Finally the experiments are re-analysed and new biological interpretation results are compared with the ones presented in the earlier analysis.

P25. Gene filtering to improve sensitivity in microarray data analysis

Stefano Calza[1,2], Yudi Pawitan[1]

[1] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

[2] *Section of Medical Statistics and Biometry, Department of Biotechnologies and Biomedical Sciences, University of Brescia, Italy*

Background: The Affymetrix GeneChips platforms have become widely used for comparing the expression of tens thousands of genes along sperimental or clinical conditions. Recent chip arrays can hold more than 50k probesets, representing almost 40,000 genes. Such an extrem abundancies of data raises many practical and theoretical problems, in terms of false discovery rate and sensitivity of the testing procedure. Nevertheless most of the measured signal is basically noise, related to non differetial expression as well as unspecific binding. An ad hoc procedure that selectively eliminate uninformative features may imporve dramatically the ability of detect real biological signal.

Our proposed method for **Filtering Likely Uninformative Sets of Hybridizations (FLUSH)** is based on robust linear models at the probe level, accounting for probes and an arrays effect. Probesets were described in terms of the Array Effect and of the standard error of the residuals. In non-normalized probesets, the Array effect will carry both the Lack of Normalization effect and the eventual Biological effect. By means of a Quantile regression is possible to a filtering criteria, based on some user-defined quantile, able to exclude features with a high residual standard error and/or low between arrays variability. The performance of the filtering procedure were evaluated on a public available dataset, with controlled spiked-in genes, in terms of bias reduction of the estimated FDR and increase of the sensitivity, as weel as on a real datataset.

Results & Conclusions: Our methods seems able to increase sostantially the sensitivity of the differentially expressed genes detection, with a less biased FDR estimation.

P26. Pathway analysis by integrating array data and regulatory motifs

Yingchun Liu[1], Markus Ringnér[1]

Department of Theoretical Physics, Lund University, Sweden

Microarrays have been used to identify gene signatures for many biological systems. It would be of importance to look for characteristic pathway activities for such gene signatures. In particular, for microarray data from experiments designed to investigate one specific pathway, it would be interesting to analyze crosstalk with other pathways. Such studies would serve as a starting point to explore dense networks of cross talking signaling pathways.

We are developing methods to identify significant pathways in microarray data by looking for the over-representation of targets of transcription factors in specific pathways.

P27. Comprehensive Bioinformatic Analysis of HIV-1 Protease Specificity

Liwen You[1,2], Daniel Garwicz[3,4], Thorsteinn Rögnvaldsson[1]

[1]*School of Information Science, Computer and Electrical Engineering, Halmstad University, Sweden*

[2]*Computational biology and biological physics, Department of Theoretical Physics, Lund University, Sweden*

[3]*Division of Hematology and Transfusion Medicine, Department of Laboratory Medicine, Biomedical Center, Lund University, Sweden*

[4]*Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Sweden*

Inhibitors of the protease of human immunodeficiency virus type 1 (HIV-1) are today an important part of highly active antiretroviral therapy (HAART) for HIV-infected individuals and AIDS patients. However, rapidly developing viral resistance to antiretroviral therapy is an increasing problem worldwide and accurate models for predicting protease cleavage specificity are needed for a rational design of more effective protease inhibitors. We have previously analyzed the specificity of HIV-1 protease using bioinformatic machine learning methods [1]. In the present work, we have extended these studies and used a new, extensive 746 peptide dataset for analysis of the specificity of HIV-1 protease [2]. We show that the best predictor is the linear predictor with sparse orthogonal coding. Although the predictor with size and hydrophobicity property coding does not perform better, size and hydrophobicity are still important two properties for the cleavage compared to other two property combination.

References:
1. Rögnvaldsson T, You L. 2004. Why neural networks should not be used for HIV-1 protease cleavage site prediction. *Bioinformatics.* 20(11):1702-1709.

2. You L, Garwicz D, Rögnvaldsson T. 2005. Comprehensive bioinformatic analysis of the specificity of human immunodeficiency virus type 1 protease. *Journal of Virology*, 79(19): 12477-86.

P28. A mixture model approach to sample size estimation in two-sample comparative microarray experiments

Tommy S. Jørstad[1] and Herman Midelfart[1]

[1]*Department of Biology, Norwegian University of Science and Technology, Norway*

Choosing an appropriate sample size is an important step in the design of a microarray experiment. A carefully chosen sample size will help control both the number of false positive conclusions and the ability to detect true differences between the samples. Controlling certain error measures through sample size, however, requires knowledge of how the truly differentially expressed genes are distributed. Estimating this distribution from observed experimental data is a difficult problem.

We present a mixture model approach to estimating the distribution of differentially expressed genes in a two-sample comparative microarray study. The algorithm for finding model parameters is in closed form. We further show how the model can be used to estimate sample sizes that control the false discovery rate (FDR) together with another statistical measure like power or false negative rate (FNR). We have tested the method on simulated and experimental data sets, and the estimates prove to be close to their true value.

P29. Bayesian hierarchical model for correcting signal saturation

Rashi Gupta[1], Petri Auvinen[2], Andrew Thomas[1], Sangita Kulathinal[3], Elja Arjas[1]

[1]*Department of Mathematics and Statistics, University of Helsinki, Finland*

[2]*Institute of Biotechnology, University of Helsinki, Finland*

[3]*Department of Epidemiology and Health Promotion, National Public Health Institute, Finland*

Microarray experiments are commonly affected by saturated pixels. Pixel saturation occurs when the pixel intensity exceeds the scanner upper threshold of detection and the recorded pixel intensity is then truncated at 65535. Truncation of the pixel intensity causes the estimate of gene expression (i.e. intensity) to be biased as a result all higher level analysis are made on these biased gene expression estimates. We propose two methods for improving the quality of signal for cDNA microarrays by making use of several scans at varying scanner sensitivities. The first method utilizes the intensities of the pixels comprising the spot and the second method utilizes the spot summary as an input to the Bayesian hierarchical model. Both models estimates and propose the true expression of genes. The methods improve the accuracy at which intensities can be measured in all ranges and extends the dynamic range of measured gene expression at the high end. The methods are generic and can be applied to data from any organism and for imaging with any scanner. Results from various real data sets illustrate an improved precision in the estimation of the expression of genes compared to what can be achieved by applying standard methods and using only a single scan.

Participants in Stochastic Centre workshop:

Statistics in Gene and Protein Expression

10-12 May 2006 at Nya Varvet, Göteborg.

Gabriella Arne	Göteborg	gabriella.arne@llcr.med.gu.se
Eva Albertsson	Göteborg	eva.albertsson@zool.gu.se
Marina Axelson-Fisk	Göteborg	marinaa@math.chalmers.se
Sören Bak	Copenhagen	bak@kvl.dk
Anders Blomberg	Göteborg	anders.blomberg@gmm.gu.se
Stefano Calza	Stockholm	stefano.calza@unimi.it
Marija Cvijovics	Berlin	cvijovic@molgen.mpg.de
Patrik Edén	Lund	patrik@thep.lu.se
Claus Ekstrøm	Copenhagen	ekstrom@dina.kvl.dk
Bjarne Ersbøll	Copenhagen	be@imm.dtu.dk
Jane Fridlyand	San Fransisco	jfridlyand@cc.ucsf.edu
Arnoldo Frigessi	Oslo	frigessi@nr.no
Peter Gennemark	Göteborg	peterg@cs.chalmers.se
Chris Glasbey	Edinburgh	chris@bioss.sari.ac.uk
Jim Graham	Manchester	jim.graham@manchester.ac.uk
Rashi Gupta	Helsinki	gupta@mappi.helsinki.fi
Lina Gunnarsson	Göteborg	lina.gunnarsson@fysiologi.gu
Mathisca deGunst	Amsterdam	degunst@cs.vu.nl
John Gustafsson	Göteborg	johng@math.chalmers.se
Anne-Mette Hein	London	a.hein@imperial.ac.uk
Anja von Heydebreck	Berlin	anja.von.heydebreck@merck.de
Eivind Hovig	Oslo	ehovig@radium.uio.no
Peter Jagers	Göteborg	jagers@math.chalmers.se
Alexandra Jauhiainen	Göteborg	jauhiain@math.chalmers.se
Margareta Jernås	Göteborg	margareta.jernas@medic.gu.se
Tommy Jørstad	Trondheim	tommy.jorstad@bio.ntnu.no
Graham Kemp	Göteborg	kemp@cs.chalmers.se
Jukka Kohonen	Helsinki	kohonen@cs.helsinki.fi
Erik Kristiansson	Göteborg	erikkr@math.chalmers.se
Morten Krogh	Lund	morten.krogh@thep.lu.se
Kim Kultima	Uppsala	kim.kultima@farmbio.uu.se
Darima Lamazhapova	Göteborg	lamazhapovadarima@llcr.med.gu.se
Mette Langaas	Trondheim	mette.langaas@math.ntnu.no
Jens Ledet Jensen	Aarhus	jlj@imf.au.dk
Alexandra Lewin	London	a.m.lewin@imperial.ac.uk
Yinchung Liu	Lund	yingchun.liu@thep.lu.se
Ted Lystig	Göteborg	theodore.lystig@astrazeneca.com
Ingrid Lönnstedt	Uppsala	ingrid@math.uu.se
Claus-Dieter Mayer	Edinburgh	claus@bioss.ac.uk
Marc Morant	Copenhagen	marmo@kvl.dk

Petter Mostad	Oslo	p.f.mostad@medisin.uio.no
Olle Nerman	Göteborg	nerman@math.chalmers.se
Hedvig Norlén	Stockholm	norlen@math.su.se
Louise Olofsson	Göteborg	louise.olofsson@medic.gu.se
Johan Palmfeldt	Göteborg	johan.palmfeldt@gmm.gu.se
Yudi Pawitan	Stockholm	yudi.pawitan@meb.ki.se
Carsten Peterson	Lund	carsten@thep.lu.se
Alexander Ploner	Stockholm	alexander.ploner@meb.ki.se
Sylvia Richardson	London	sylvia.richardson@imperial.ac.uk
Markus Ringnér	Lund	markus@thep.lu.se
Mats Rudemo	Göteborg	rudemo@math.chalmers.se
Janeli Sarv	Göteborg	sarv@math.chalmers.se
Anders Sjögren	Göteborg	anders.sjogren@math.chalmers.se
Gordon Smyth	Melbourne	smyth@wehi.edu.au
Rolf Sundberg	Stockholm	rolfs@math.su.se
Per Sunnerhagen	Göteborg	molps@lundberg.gu.se
Ziad Taib	Göteborg	ziad.taib@astrazeneca.com
Natalie Thorne	Cambridge	npt22@cam.ac.uk
Michael Thorsen	Göteborg	michael.thorsen@gmm.gu.se
Wessel van Wieringen	Amsterdam	wvanwie@few.vu.nl
Liwen You	Lund	liwen@thep.lu.se
Magnus Åstrand	Göteborg	magnus.astrand@astrazeneca.com
Lisa Öberg	Göteborg	lisa.oberg@astrazeneca.com