

# *Symposium on* **Data Analysis in Transcriptional Studies**

9:15 h to 13:15 h  
Friday, 15 December 2006  
Buchleitner seminar room, TP3,  
DKFZ, Im Neuenheimer Feld **580**

9:15-9:30 Introduction

9:30-10:00 **Martin Kuiper**  
*Ghent University, Belgium*

EMERALD: a FP6 Coordination Action to promote development of standards and quality metrics for microarray technology.

10:00-10:30 **Alvis Brazma**  
*European Bioinformatics Institute, Hinxton,  
Cambridge, England*

Intelligence in ArrayExpress, or how do I find data that tells me something about the expression of my favourite gene.

- 10:30-11:00 **Joaquin Dopazo**  
*Principe Felipe Research Center, Valencia, Spain*  
Towards a systems biology-based approach to the interpretation of genome-scale experiments.
- 11:00-11:15 coffee break
- 11:15-11:45 **Reinhard Guthke**  
*Leibniz-Institute for Natural Product Research and Infection Biology, Jena*  
Network inference from gene expression data.
- 11:45-12:15 **Stefan Winter**  
*Stuttgart University*  
Regression based survival analysis of DNA microarray data.
- 12:15-12:45 **Eugene van Someren**  
*Delft University of Technology, The Netherlands*  
Integrated Network Analysis unravels the osteoblast differentiation pathway.
- 12:45-13:15 **Christine Steinhoff**  
*Max-Planck-Institute for Molecular Genetics, Berlin*  
Impact of DNA methylation on gene expression and regulation.

## **EMERALD: a FP6 Coordination Action to promote development of standards and quality metrics for microarray technology**

**Martin Kuiper**, Arne Sandvik, Alvis Brazma, Carole Foy, Joaquin Dopazo, Wolfgang Philipp, Laszlo Puskas, Ulf Landegren

Recently, the Coordination Action EMERALD was launched, running until November 2009. The objective of this CA is to establish and disseminate quality metrics (QC), microarray standards and best laboratory practices (QA) throughout the European microarray community. The EMERALD core consortium comprises a number of different stakeholders of microarray technology. This includes the main research and innovation operators involved in the development of microarray standards and quality metrics (EBI, LGC, IRMM), the stakeholders in the data production process (core facilities, companies, technology innovators: NTNU, BRC, UU), and experts in information extraction and data modelling (NTNU, VIB, CIPF). Some partners also play important roles in MGED or ERCC. The 'grassroots movement' that became the Microarray Gene Expression Data Society (MAGE Society) has established guidelines for experiment description (MIAME: Minimum Information About a Microarray Experiment) and a description of a structured data exchange model (MAGE-ML). MGED initiatives have predominantly been focused on data context, but here we want to focus more on data content. Microarray data will increasingly become essential for data driven model building in systems biology approaches. Quality and coherence of microarray data compendia (e.g. in ArrayExpress) are major determinants for information extraction and model building. This Coordination Action is designed to structure and organise these efforts at a European scale, this in close association with MGED and ERCC.

## **Intelligence in ArrayExpress, or how do I find data that tells me something about the expression of my favorite gene**

**Alvis Brazma**  
EMBL-RBI

We have accumulated data from over 50 000 microarray hybridisations and over one and a half million expression profiles in ArrayExpress public database of microarray experiments. It is not a trivial question, how given a gene or a group of genes, one can find microarray experiments that are informative about the expression of this particular gene or genes. We have developed algorithms that are based on linear models and on mutual information, which help to answer these questions. In the talk I will show the new functionality of ArrayExpress, discuss the research issues around the search the new ArrayExpress search engines, as well as some applications to biomedical research.

See [www.ebi.ac.uk/arrayexpress](http://www.ebi.ac.uk/arrayexpress).

## **Towards a systems biology-based approach to the interpretation of genome-scale experiments**

**Joaquin Dopazo**

Bioinformatics Department, CIPF, Valencia, Spain.

With the popularisation of high-throughput techniques the need for procedures that help in the biological interpretation of the deluge of data produced has increased enormously. This interpretation has initially been made on a pre-selection of "interesting" genes, based on the experimental values, followed by the study of their functional properties. The study of functional properties include different ways of checking the over-representation of terms with functional meaning (e.g. gene ontology, pathways, etc.) Unfortunately, these approaches, which inherit some pre-genomic paradigms, basically ignore the coordinated behaviours of the genes. Recently, new procedures inspired in systems biology criteria have started to be developed (Al-Shahrour et al. 2005a; Mootha et al. 2003). These new procedures represent a conceptual change in the hypothesis to test given that they aim to directly test the behaviour of blocks of functionally related genes, instead of focusing on single genes. This function-centric view is gaining credence and popularity within the scientific community as new tools to undertake such analyses are becoming available (Al-Shahrour et al. 2005b).

Al-Shahrour, F., R. Diaz-Uriarte, and J. Dopazo. 2005a. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* **21**: 2988-2993.

Al-Shahrour, F., P. Minguéz, J.M. Vaquerizas, L. Conde, and J. Dopazo. 2005b. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res* **33**: W460-464.

Mootha, V.K., C.M. Lindgren, K.F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M.J. Daly, N. Patterson, J.P. Mesirov, T.R. Golub, P. Tamayo, B. Spiegelman, E.S. Lander, J.N. Hirschhorn, D. Altshuler, and L.C. Groop. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**: 267-273.

## Network Inference from Gene Expression Data

Reinhard Guthke

Leibniz Institute for Natural Product Research and Infection Biology – Hans Knoell  
Institute (HKI), D-07745 Jena, Beutenbergst. 11a

Discovering and understanding the complex molecular interactions that make up living organisms is one of the most interesting and challenging problems of modern molecular biology, systems biology and bioinformatics. We present an optimized reverse engineering strategy aimed at a reconstruction of dynamic gene interaction networks. The proposed approach is based on both, microarray data and available biological knowledge. The main kinetics of the response after a perturbation is identified by fuzzy clustering of gene expression profiles (time series). The number of clusters is optimized using various evaluation criteria. For each cluster a representative gene with a high fuzzy-membership is chosen in accordance with available physiological knowledge. Then hypothetical network structures are identified by seeking systems of ordinary differential equations, whose simulated kinetics could fit the gene expression profiles of the cluster-representative genes. Resampling methods are applied to validate the obtained dynamic network models and to improve their generalization strength. This novel reverse engineering algorithm was implemented in MATLAB [1] and applied to different sets of DNA microarray data, such as characterizing the stress response of *Escherichia coli* after induction of recombinant protein synthesis [2], the immune response after infection of human blood cells by pathogenic *E. coli* [3], the response of hepatocytes after stimulation by LiCl to study the regulation of wnt/ $\beta$ -catenine pathway [5] and the response of the human-pathogenic fungus *Aspergillus fumigatus* after temperature shift [5]. The first application was already continued by integrated analysis of transcriptome and proteome data [6]. The latter application will be presented more in detail. Time series data, i.e. expression profiles of 1926 differentially expressed genes of *A. fumigatus* [7] were clustered by fuzzy c-means. The number of clusters was optimized using a set of optimization criteria. From each cluster a representative gene was selected by text mining in the gene descriptions and evaluating gene ontology terms. The expression profiles of these genes were simulated by a differential equation system, whose structure and parameters were optimized minimizing both the number of non-vanishing parameters and the mean square error of model fit to the microarray data.

- [1] Toepfer S, Guthke R, Driesch D, Woetzel D, Pfaff M (2007) The NetGenerator Algorithm: Reconstruction of Gene Regulatory Networks. *Lecture Notes in Bioinformatics*, 4366, in press.
- [2] Schmidt-Heck W, Guthke R, Toepfer S, Reischer H, Dürrschmid K, Bayer K (2004): Reverse Engineering of the Stress Response during Expression of a Recombinant Protein. Proc. EUNITE Symp. 10.-12. June 2004, Verlag Mainz, Aachen, pp. 407-412.
- [3] Guthke R, Möller U, Hoffmann M, Thies F, Töpfer S (2005): Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21, 1626-1634.
- [4] Zellmer S, Schmidt-Heck W, Gaunitz F, Baldysiak-Figiel A, Guthke R, Gebhardt R (2005): Dynamic Network Reconstruction from Gene Expression Data Describing the Effect of LiCl Stimulation on Hepatocytes. *Journal of Integrative Bioinformatics*, 0015, *Online Journal*: [http://journal.imbio.de/index.php?paper\\_id=15](http://journal.imbio.de/index.php?paper_id=15)
- [5] Guthke R, Kniemeyer O, Albrecht D, Brakhage AA, Möller U (2007): Discovery of Gene Regulatory Networks in *Aspergillus fumigatus*, *Lecture Notes in Bioinformatics*, 4366, in press.
- [6] Schmidt-Heck W, Dürrschmid K, Reischer H, Hrebicek T, Rizzi A, Bayer K, Guthke R (2006): Analysis of Transcriptome and Proteome Time Series Data from Recombinant *Escherichia coli* Cultivations, submitted
- [7] Nierman, W.C. et al. (2005): Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*, 438, 1151-1156.

## **Regression based survival analysis of DNA microarray data**

**Stefan Winter**

Fachbereich Mathematik, University of Stuttgart

Survival prognosis, e.g. estimation of time to death or relapse, is an important task in the analysis of clinical data sets. In this censored regression problem, we show how one can define estimates of the mean survival time, the conditional variance, and the conditional survival function. In biomedical research, DNA microarrays provide up to several thousand possible genetic markers, which can enhance predictions solely based on clinical and histological covariables. On one hand, just a small percentage of these genes are considered to have real predictive power. On the other hand, due to the "curse of dimensionality", regression analysis in the presence of high-dimensional covariates is especially difficult. One possibility to overwhelm this is to agglomerate several correlated genetic markers in one predictive covariable. On the basis of such covariables, we will illustrate our estimates by applying them to simulated and real (human cancer) data sets.

Dippon, J. and Winter S. (2006). Smoothing spline regression estimates for randomly right censored data. Preprint 2006-006, Fachbereich Mathematik, Universität Stuttgart.

## **Integrated Network Analysis unravels the osteoblast differentiation pathway**

**Eugene van Someren**

Delft University of Technology

The aim of these studies is to identify the interactions between genes that drive the process of osteogenesis. The ultimate goal is to discover drug targets for osteoporosis. In order to understand any developmental process and discover possible points of intervention, it is imperative to unravel the underlying genetic interaction network that drives this process. For this purpose, different time-series microarray measurements are available with difference in cell-line, micro-array platform, no. samples and sampling scheme.

We investigated a large number of different reverse engineering schemes to discover genetic regulation from time-series microarray data set during osteoblast development. One promising approach estimates the interaction network by solving a linear model using simultaneous shrinking of the least absolute weights and the prediction error. This approach effectively solves the problem of having a limited amount of arrays by focussing on finding the structure of the network.

Microarray data alone is not sufficient to adequately solve this task and in this presentation we will show how our analysis benefits from the integrated use of different bioinformatics tools (e.g. genetic network modeling and promotor analysis) and data-sources such as different micro-array datasets, sequence information and different functional annotations (e.g. literature networks and GO enrichment).

- E.P. van Someren, B.L.T. Vaes, W.T. Steegenga, A.M. Sijbers, K.J. Dechering and M.J.T. Reinders, Least Absolute Regression Network Analysis of the Murine Osteoblast Differentiation Network, *Bioinformatics*, 22, 4, pp. 477-484, 2006.
- Bart L.T. Vaes, Patricia Ducey, Anneke M. Sijbers, José M. A. Hendriks, Eugene P. van Someren, Nanning G. de Jong, Edwin R. van den Heuvel, Wiebe Olijve, Everardus J.J. van Zoelen and Koen J. Dechering, Micorarray analysis on runx2-deficient mouse embryos reveals novel runx2 functions and target genes during intramembranous and endochondral bone formation, *Bone*, 39(4):724-38, 2006.

## **Impact of DNA Methylation on Gene Expression and Regulation**

### **Christine Steinhoff**

Max Planck Institute for Molecular Genetics, Berlin, Germany.

Epigenetic modifications play a crucial role in gene regulatory processes. Among these, DNA methylation is one major epigenetic modification that directly affects gene expression. DNA methylation patterns are heritable and meanwhile it is well known that they largely differ comparing various developmental stages and tissues. DNA methylation is involved for example in transcriptional regulation of genes and the establishment of tissue type specific patterns, silencing of retroelements, control of imprinted genes, cancer development and progression. Thus, in order understand regulatory processes approaches on how DNA methylation affects directly gene expression and further on cellular regulatory processes are necessary. Therefore, we studied genomewide tissue specific gene expression in human and mouse and derived a score that accounts for local and tissue specific variance in expression. Using this score we define candidate regions where DNA methylation might play a regulatory role and performed a detailed analysis of these regions regarding the correlation with a number of sequence features. Furthermore we examined the DNA methylation patterns in specific candidate regions.