

Inference of regulatory relationships in human urothelial cancer

Matthias Böck and Constanze Schmitt

TU München, Lehrstuhl I12 Bioinformatik

RECESS Retreat
Schloss Hohenkammer
June 22, 2010

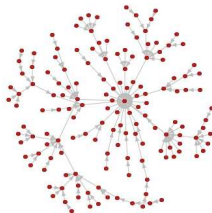
Outline

- 1 Introduction
 - Research plan
 - Data
- 2 Quality check
- 3 Steady state data (cancer)
- 4 Time series (NHU)
 - Introduction and goals of the analysis
 - R tools for analysing time series
 - Preliminary results with maSigPro
- 5 Outlook

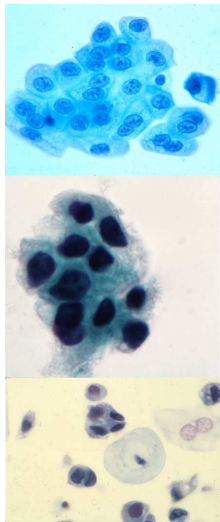
Research plan

Gene regulatory network inference

- Integration of different data types (time series, steady states, etc.)
- Build preliminary model on available data with possibility to feed in more data types (York)
- Focus on qualitative description to make rather general but hopefully more stable models
- Constraint-based learning (e.g., using a logic programming formalism) for inferring and confining network substructures (e.g. alternative paths, inferring of edge directions given partial structure)



Data basis for our projects

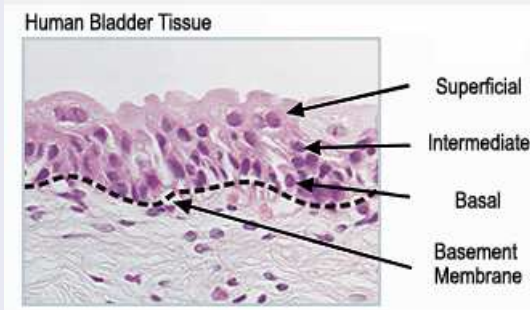


INSIGH2

- Collaboration between TUM, Institut Curie and the University of York
- Unravel the coupled relationships between tumorigenesis and differentiation in a human tissue
 - Use transcriptomic as well as proteomic data to study human urothelium
 - Focus on the dedifferentiation of cells rather than proliferation
 - Integrate available data sets into a shared database
 - Analyse the experimental data sets and infer regulatory networks

Biological background: urothelium

- Very specialized type of epithelium; capacity to self-renew; normally mitotically quiescent
- 3 cell layers, protected by uroplakins (UPKs) on apical surface; basal to differentiated cells;
- Stretchable permeability barrier
- Differentiation/proliferation as wound response



Data up to now

Patient data (York)

Samples of healthy patients → NHU in vitro cell lines

Aim: induce differentiation and proliferation

Troglitazone model

- Troglitazone: activator of PPAR γ -receptors (TFs)
- Tissue function lost
- Study urothelial differentiation

ABS/Ca²⁺ model

- Biomimetic tissue
- Low Ca²⁺ concentration leads epithelial cells to proliferate more quickly
- Study wound closure → urothelial differentiation and proliferation

Cancer data

Tumor samples (Institut Curie)

57 bladder carcinoma samples;

Compared healthy and tumor expression correlation and determine DNA copy-number independent deregulation

→ epigenetic mechanisms

Additional data (Institut Curie)

80 tumor patient samples

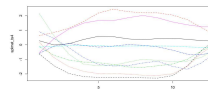
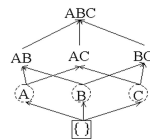
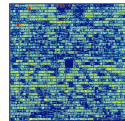
Data integration

Database

- Integrate available annotations
- Combine with cancer data from Curie Institute
- Integrate additional publicly available data sets

Analysis

- Clustering, correlation and pattern search
 - Involved genes
 - Different cell stages
 - Co-expressed genes
 - ...
- Causality from time series?
- Integrate prior knowledge like TFs, existing publications, ...



Images of raw data

P2307R2U95A

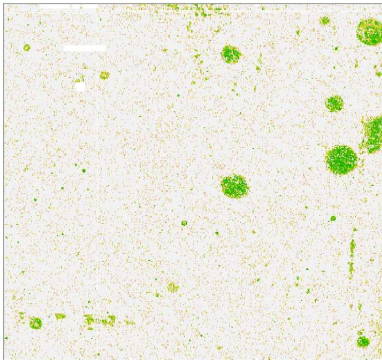


Figure: drops of water?

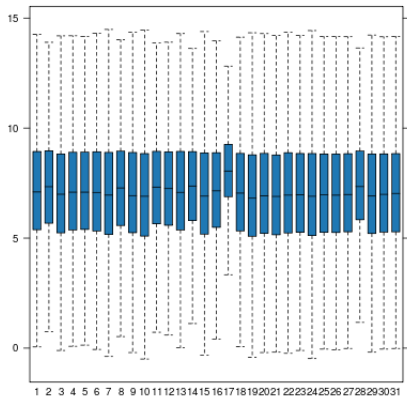
P1512-6R1V2



Figure: good quality

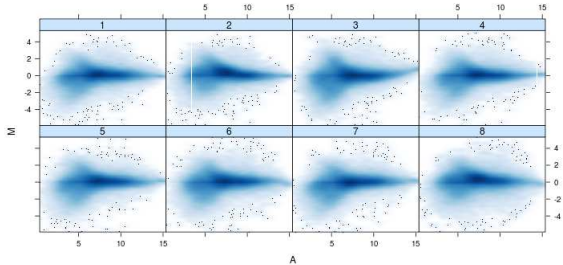
arrayQualityMetrics (Bioconductor)

- Boxplots
- MA-plots
- Heatmap



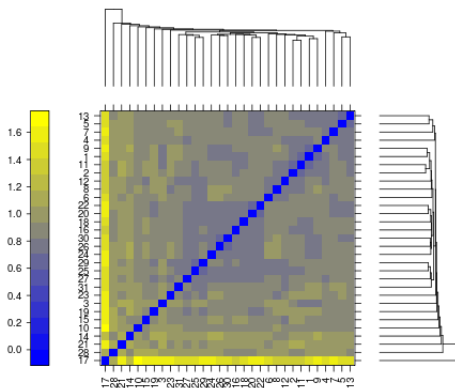
arrayQualityMetrics (Bioconductor)

- Boxplots
- MA-plots
- Heatmap



arrayQualityMetrics (Bioconductor)

- Boxplots
- MA-plots
- Heatmap



Possible difficulties with the data

Batch-effects

- Non-biological experimental variation
- Chips of one batch are generally more similar to each other than to chips in other batches
- Causes: amount of amplification agent used, time, session
- Data from different batches non-comparable

Possible remedies

ComBat tool for adjusting batch effects, based on empirical Bayes model

Assumption: batch effects are similar across genes (i.e. increased expression, higher variability,...)

⇒ Pool information across all genes to shrink batch effect parameter estimates toward a common mean

Normalization

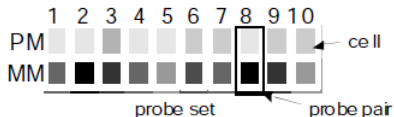


Figure: Affymetrix GeneChip Array design

Normalization is done to make to scale different arrays and thus make them comparable

- RMA (robust multichip average)
- MAS5 (standard affy)
- VSN (genes with little variance)

Differential expression analysis

Linear Models, T-test

- Linear model is fitted for each gene
- $\text{Diff} = \alpha_1 * \text{probe}_1 + \dots + \alpha_n * \text{probe}_n$

Empirical bayes

eBayes moderation of standard errors
Standard errors across genes shrunk to common value

Ranking and multiple testing correction

Multiple testing problem: simultaneously considering a set of statistical inferences leads to “false positives”

Example

Test 50 000 probes for differential expression with type 1 error limited to $\alpha = 0.05$ (2500 falsely classified genes expected)

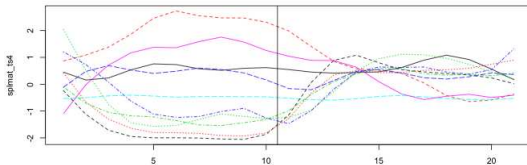
Correction: readjust α e.g.: FDR

Preliminary results on cancer data

Sets of top-differentially expressed genes

- Stratification after different confounders (tumor stages)
- Differential expression analysis with respect to NHU samples
- Comparison to KEGG bladder cancer pathway (42 genes)
- No rediscovery; but found genes potentially tumor-associated (e.g. cytokines inhibiting tumors, metalloproteins implicated in metastasis)
- “Cancer is a heterogeneous landscape”

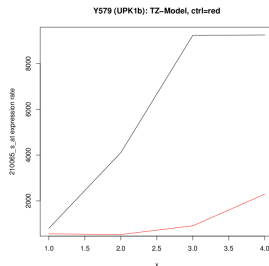
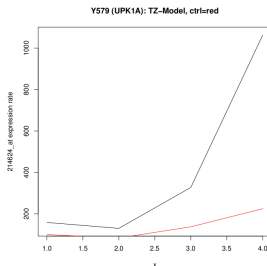
Short introduction to time series



Time series

- A sequence of data points, measured typically at successive times spaced at uniform time intervals
- In biology usually short series with about 4 or 5 points in time
- In our case 4 and 2 different cell lines treated with 2 different methods and 2 untreated controls ($4 \times 2 \times 2 \times 2 = 32$ Chips)
- Reactions (temporal and longitudinal) of the cell lines should reflect dependencies between the genes

Short introduction to time series



Time series

- Usually 3-5 replicates for each experiment. In our case 2 and control cell lines are only to a certain extent similar
- Differences between ABS and TZ model as well as cell lines
- Analyse behavior of Uroplakins (UPK1a/b, UPK2, UPK3a/b) and some other factors which are supposed to be marker for differentiation

⇒ Find gene cluster, regulators of these genes, define developmental stages...

R tools for analysing time series

Bioconductor tools

- timecourse package
- maSigPro package
- lmFit (define own regression model)

Create a design matrix describing the models, then apply a (mostly linear) regression fit on that data and use a suitable ranking for the genes or define a specific contrast matrix

maSigPro - microarray Significant Profiles

- Define the regression model (in our case a quadratic regression model)
- Do a regression fit for each gene \Rightarrow find significant genes
- Stepwise regression on the variables \Rightarrow find significant profile differences between experimental groups
- Get significant genes p.e. with an expression profile significantly different from a 0 profile

Quadratic regression model

$$y_{ijr} = \beta_0 + \beta_1 D_{(UT)ijr} + \beta_2 D_{(LO)ijr} + \beta_3 D_{(ME)ijr} + \beta_4 D_{(HI)ijr} \\ + \beta_5 T_{ijr} + \beta_6 D_{(UT)ijr} \times T_{ijr} + \beta_7 D_{(LO)ijr} \times T_{ijr} \\ + \beta_8 D_{(ME)ijr} \times T_{ijr} + \beta_9 D_{(HI)ijr} \times T_{ijr} + \beta_{10} T_{ijr}^2 \\ + \beta_{11} D_{(UT)ijr} \times T_{ijr}^2 + \beta_{12} D_{(LO)ijr} \times T_{ijr}^2 + \beta_{13} D_{(ME)ijr} \\ \times T_{ijr}^2 + \beta_{14} D_{(HI)ijr} \times T_{ijr}^2 + \varepsilon_{ijr}.$$

- 4 dummy variables D
- y expression value; T time; β regression coefficients
- i experimental groups; j time points, r replicates

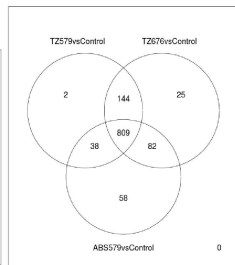
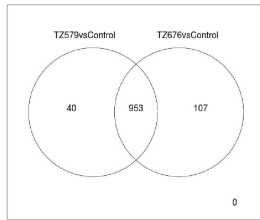
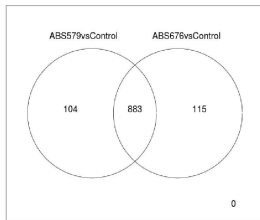
Preliminary results with maSigPro - list of significant genes

- 54.613 probe sets in total (about 38.000 genes)

⇒ Apply non-specific filtering to reduce noise and dimensionality of the data
 ⇒ Intensity filter for genes which are above 100 in at least 0.25 of all states

- 36.388 remaining probe sets from which 1158 sets (811 genes) have a significant profile:

Experiments	TZ-Y579	TZ-Y676	ABS-Y579	ABS-Y676
Significant	993	1060	987	998



Preliminary results with maSigPro - cluster of significant genes

- 9 clusters are created with hclust (similar profiles)
- UPK3a, UPK2, UPK1a are in the same cluster (8), UPK3b and UPK1b are in another one (5)

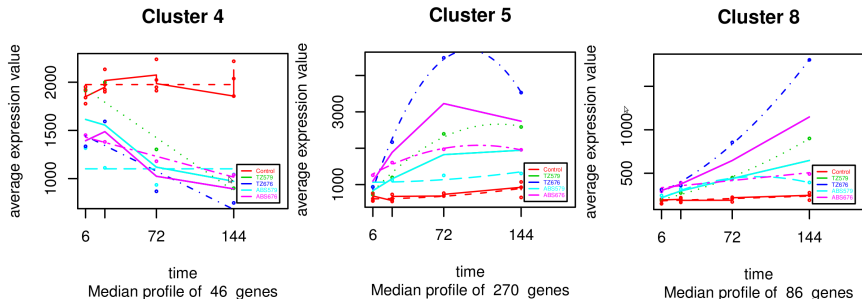


Figure: Group expression profiles

Uroplakins as differentiation markers

- Urothelium-specific markers of terminal urothelial cytodifferentiation
- Genes: UPK1a, UPK1b, UPK2, UPK3a, UPK3b (1a and 2 may be urothelium specific)
- Plaques of the AUM (asymmetric unit membrane)

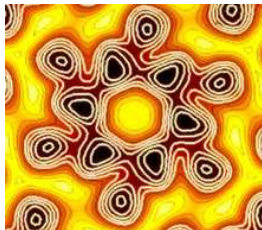


Figure: Uroplakin by electron microscopy
(<http://urology.med.nyu.edu/research/center-of-excellence/translational-bladder-cancer-research>)

Combined results

- Most UPKs show high log-fold change
- Best ranks of UPK3A and 1A in tumor stage Ta (cancer cells only in bladder lining, low contamination)
- UPK2 appears only in stage T4
- Relation to differentiation loss
- \Rightarrow 3A,1A are also in one cluster in ts analysis \Rightarrow one regulation mechanism?

Further steps in our analysis

- Deal with high dimensionality of the data
- More knowledge has to be integrated into our predictions
 - TF analysis \Rightarrow Prof. Gelfand
 - Information on alternative splicing \Rightarrow Prof. Gelfand
 - Database knowledge (EMBL, String,...)

Data integration with ProbLog

- Use weight-of-evidence (given or inferred) approach to build a background knowledgebase
- A program consists of a set of definite clauses
- Each clause is labeled with a probability
- These rules form the basis for the database and are trained with given facts
- Queries can be sent to the final knowledgebase \Rightarrow answer with probability

Knowledgebase

Clauses:

`likes(X,Y):- friendof(X,Y).`

Prob. clauses:

`0.8: likes(X,Y):- friendof(X,Y),likes(Z,Y).`

Facts:

`0.5: friendof(john, mary).`

Apply on networks

Clauses:

`path(X,Y):- dir-edge(X,Y), ... , ...`

`coregulated(X,Y) :- dir-edge(Z,X), dir-edge(Z,Y), ...`

`interact(X,Y) :- dir-edge(X,Y) ; dir-edge(Y,X).`

Facts:

`0.5: coregulated(X,Y).`

\Rightarrow Optimize for `?: dir-edge(X,Y)`

Data integration with ProbLog

- Use weight-of-evidence (given or inferred) approach to build a background knowledgebase
- A program consists of a set of definite clauses
- Each clause is labeled with a probability
- These rules form the basis for the database and are trained with given facts
- Queries can be sent to the final knowledgebase \Rightarrow answer with probability

Knowledgebase

Clauses:

likes(X,Y):- friendof(X,Y).

Prob. clauses:

0.8: likes(X,Y):- friendof(X,Y),likes(Z,Y).

Facts:

0.5: friendof(john, mary).

Apply on networks

Clauses:

path(X,Y):- dir-edge(X,Y), ... , ...

coregulated(X,Y) :- dir-edge(Z,X), dir-edge(Z,Y), ...

interact(X,Y) :- dir-edge(X,Y) ; dir-edge(Y,X).

Facts:

0.5: coregulated(X,Y).

\Rightarrow Optimize for ?: dir-edge(X,Y)

SOMs

- Related to neural networks; iterative unsupervised learning process
- Discretized representation of input while preserving topological properties (stepwise adaptation to topology)
- Neighbourhood: nodes should “respond” to similar stimuli

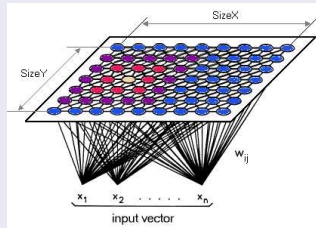


Figure: A Kohonen map with quadratic grid

Thank you.