

# Data-based Extraction of Hypotheses about Gene Regulatory Networks in Liver Cells

Wolfgang Schmidt-Heck<sup>1</sup>, Sebastian Zellmer<sup>2</sup>, Frank Gaunitz<sup>3</sup>, Rolf Gebhardt<sup>2</sup>, Reinhard Guthke<sup>1</sup>

<sup>1</sup>Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knoell-Institute, Beutenbergstr. 11a, D-07745 Jena, Germany, Phone: +49-3641-656820, Fax: +49-656825, email: {wolfgang.schmidt-heck, reinhard.guthke}@hki-jena.de

<sup>2</sup>Institute for Biochemistry, Medical Faculty, University Leipzig, Liebigstrasse 16, D-04103 Leipzig, Germany, Phone: +49-0341-9722167, -9722100, Fax: +49 341-9722109 email: {sebastian.zellmer, rgebhardt}@medizin.uni-leipzig.de

<sup>3</sup>Interdisciplinary Centre for Clinical Research (IZKF) Leipzig, Inselstrasse 22, D-04103 Leipzig, Germany, Phone: +49-341-97 22153, Fax: +49-341-9722109, email: frank.gaunitz@medizin.uni-leipzig.de

**ABSTRACT:** Monitoring of gene expression was used to generate hypotheses about gene regulatory networks in the liver. Due to the large number of possible interaction partners the selection of relevant nodes of the network was the crucial step of data analysis. This selection was performed by filtering differentially expressed genes at different sampling time points. The arcs of the network were revealed by the NetGenerator algorithm that minimized both, the model fit error and the number of non zero model parameters under the restriction of available biological knowledge. The reverse engineering approach is applied to elucidate the signaling pathway in liver cells which is responsible for spatial organization of the liver, in particular the different physiological patterns in periportal and pericentrally located hepatocytes.

**KEYWORDS:** Systems Biology; Hepatozytes; Clustering; Network Reconstruction; Bootstrapping

## INTRODUCTION

The liver is the main organ of intermediate metabolism with regulatory, detoxifying and metabolic functions. For example, components of food, such as carbohydrates, proteins and lipids, are metabolized in the liver to generate energy (ATP) and are broken down into smaller molecules, such as amino acids, ammonia, urea, CO<sub>2</sub> and water. Other physiological important proteins, such as albumin, are synthesized in the liver. Xenobiotics, such as drugs, alcohol or coffee are detoxified. The liver mainly consists of hepatocytes (80 % of the liver volume). The hepatocytes are arranged in a spatial organization, which results in a periportal and a pericentral location, i.e. cells located near to the portal vein or to the central vein, respectively. Depending on the location of the individual hepatocyte within the liver it differs in its enzyme/protein expression pattern and metabolic capacity [1]. For instance, glycolysis ( degradation of glucose) takes place at a higher rate in pericentral hepatocytes, whereas gluconeogenesis (formation of glucose), is higher in periportal cells. Some enzymatic activities like glutamine synthetase are located only in a small subset of pericentral hepatocytes surrounding the veins. The spatial organization of the liver cells allows an efficient adaptation of liver metabolism to the different nutritional requirements of the whole organism at different metabolic states. A systems biological approach can help to understand the regulation of liver zonation.

Experimental evidence exists that the Wnt/ $\beta$ -catenin signaling pathway plays an important role in the zonation of liver parenchyma [2]. To elucidate the role of this signaling pathway the gene expression pattern of hepatocytes was monitored and analyzed by reverse engineering techniques after stimulation by LiCl. LiCl inhibits the glycogen synthase kinase-3 $\beta$  (GSK3 $\beta$ ), a key regulator of the Wnt/ $\beta$ -catenin pathway and induces the glutamine synthetase (GS), a marker protein of pericentral hepatocytes [3]. The systems biological approach of this study generated hypothetical regulatory pathways, which can explain the unique expression of GS in pericentral hepatocytes. In addition, the methodology of data-based modeling and the robustness of the obtained results are discussed.

The Human Genome Project opened the door for a holistic view in molecular medicine and to personalized medicine. Creating comprehensive models that can predict cellular behaviors are one of the major goals of systems biology [4, 5]. This requires the integration of experimental and computational approaches. The biological information becomes available by tools of high-throughput quantitative measurements(e.g. DNA sequencing in genomics, DNA arrays in

transcriptomics, gel electrophoresis and mass spectrometric techniques in proteomics, etc. Computer science, mathematics and statistics are employed to analyze and integrate biological information for deciphering complex biological systems. The aim is to develop models that allow the description and simulation of these processes on the basis of experimental data “in silico”. An understanding of gene regulatory networks and signaling pathways can be complicated by complex experimental set-ups for the identification of protein interactions and quantification. Mathematical modeling, simulation and data generation is an iterative process by which a hypothesized model is validated through indirect measurements. In turn the model and simulations can guide further experimental design. In this paper microarray data were used to elucidate the structure and dynamics of a gene regulatory network. Representative genes describing the nodes of the network were selected by data filtering and clustering techniques together with exploitation of biological knowledge. The network structure was identified by a heuristic structure optimization algorithm.

## MATERIAL AND METHODS

The gene expression data of hepatocytes with and without stimulation of LiCl were monitored using Affymetrix GeneChip Hg-U133A oligonucleotide arrays, which were pre-processed using the ‘affyPLM’ packages of the Bioconductor Software [6, 7]. The data were plotted as logarithmized ( $\log_2$ ) ratios (“log-ratio”) of the expression intensity of 22283 probesets (in the following called ‘genes’; probesets represent genes or ESTs) at 6 time points  $t$  ( $t = 0, 2, 4, 8, 12, 24$  h). Data obtained from cells without LiCl stimulation were subtracted.

Profiles of genes differentially expressed by a fold larger than two, i.e. with an absolute value of a log-ratio greater one in one or more samples were clustered according to the temporal allocation of the extremum. Networks of interactions between genes representing gene clusters were obtained and optimized using a heuristic Network Generation Method recently published [8] and implemented in MATLAB.

## RESULTS

### CLUSTERING OF GENE EXPRESSION PROFILES

A subset of 662 genes was found to be differentially expressed by a fold greater two in one or more samples, i.e. the absolute value of log-ratio surpasses 1 at least at one time point. The profiles of the differentially expressed genes were clustered according to the temporal allocation of the minimum or maximum. As the result 10 different clusters were identified as shown in Figure 1. Table I shows the number of genes ( $N$ ) belonging to a cluster as well as a representative gene.

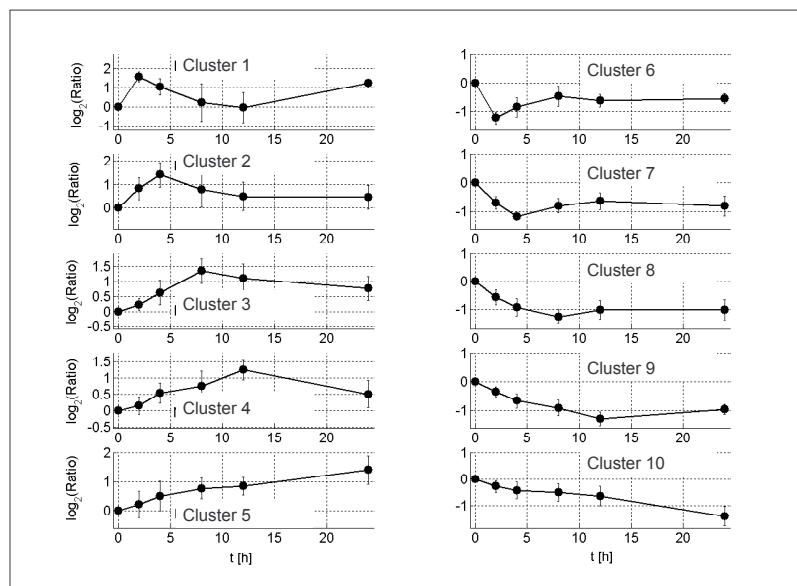


Fig. 1: Result of the clustering according to the temporal allocation of the minimum or maximum: Mean normalized gene expression profiles with standard deviation averaged over the  $N$  genes for the respective cluster (Table 1)

c	T	E	N	ID	Symbol	Description
1	2 h	+	2	201289_at	CYR61	cysteine-rich, angiogenic inducer, 61
2	4 h	+	10	209774_x_at	CXCL2	chemokine (C-X-C motif) ligand 2
3	8 h	+	12	204731_at	TGFBR3	transforming growth factor, beta receptor III
4	12 h	+	51	209610_s_at	SLC1A4	solute carrier family 1 (glutamate/neutral amino acid transp.)
5	24 h	+	27	202859_x_at	IL8	interleukin 8
6	2 h	-	8	203739_at	ZNF217	zinc finger protein 217
7	4 h	-	28	202948_at	IL1R1	interleukin 1 receptor, type I
8	8 h	-	23	218146_at	AD-017	glycosyltransferase AD-017
9	12 h	-	31	201092_at	RBBP7	retinoblastoma binding protein 7
10	24 h	-	524	208407_s_at	CTNND1	catenin (cadherin-associated protein), delta 1

Table I. Number ( $N$ ) of genes belonging to cluster  $c$  and characterized by the allocation time  $T$  of the extremum  $E$ . +/- denotes the maximum or minimum. In addition selected representative gene (Probeset ID, Symbol, Description) are given.

## NETWORK STRUCTURE OPTIMIZATION FOR DYNAMIC MODELLING

The response of hepatocytes after the stimulation with LiCl was simulated by dynamic models using the results of the cluster analysis. The resulting network structure (Figure 2) was optimized using the NetGenerator algorithm [8]. This reverse engineering algorithm minimized the mean square error of the model fit of the gene expression data as well as the number of non-zero parameters of the linear differential equation system. The maximum dynamic order  $R_i$  for each measured variable was set to three (allowing up to two additional delay elements per measured variable). The maximum allowed error for the model fit to the measured and pre-processed log-ratios was set to one (corresponding a minimum fold change two). A linear differential equation system (1) was developed to simulate the kinetic profiles (Figure 3). Table II shows the model parameters identified by model fit. The general model structure (1) involved 10 sub-models of the dynamic order  $R_i$ . For the expression kinetics of two of the 10 selected genes the dynamic order of 3 were used whereas for the other eight genes the dynamic order of one was found to be sufficient. Eight gene-to-gene interactions were found ( $w_{ij} > 0$ ). For five of them the gene AD-017 was the source.

$$\begin{aligned} \frac{dx_{i,1}}{dt} &= \sum_{j \in D_i} w_{i,j} \cdot x_j(t) + w_{i,i} \cdot x_{i,1}(t) + b_i \cdot u(t) \\ \frac{dx_{i,r}}{dt} &= x_{i,r-1}(t) + w_{i,i} \cdot x_{i,r}(t); \quad r = 2, \dots, R_i - 1 \\ \frac{dx_i}{dt} &= x_{i,R_i-1}(t) + w_{i,i} \cdot x_i(t) \end{aligned} \tag{1}$$

for  $i = 1, \dots, 10$ .

## MODEL VALIDATION

To validate the received network model, normally distributed noise with the mean zero and several standard deviations  $\sigma = 0.05, 0.20$  and  $0.30$  were added to the measured and pre-processed log-ratios and the model was reconstructed 1000 times. A gene-gene interaction  $w_{ij}$  or stimulus response component  $b_i$  was accepted as part of the network, if it occurred in at least 90 percent of the reconstructed models (Figure 2, Table II column ‘valid’). Four of the eight gene-to-gene interactions could not be validated.

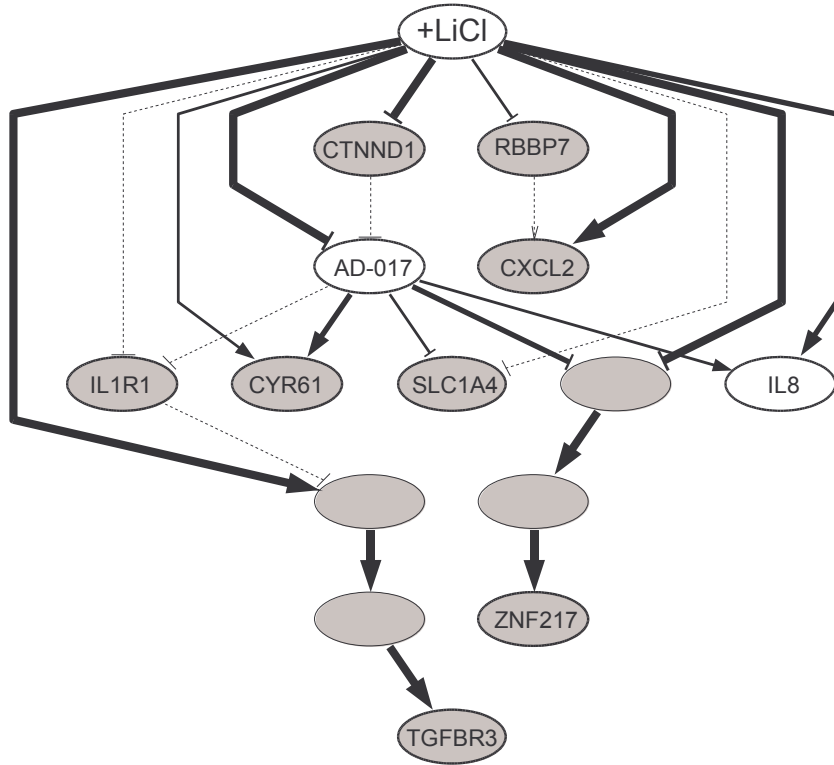


Fig. 2. Network structure of the dynamic model identified by the NetGenerator algorithm [8] using the expression profiles of the cluster representative genes (Fig. 3). The arrows represent stimuli or activations. The T-shaped links ( $\perp$ ) represent inhibitions or repressions. Thick, medium and thin lines indicate a robust gene-to-gene interaction occurring in at least 90%, 0.20 and 0.05, respectively. The dashed lines represent interactions which were not validated by bootstrap analysis. White nodes for AD-017 and IL8 denote that for these genes the self-regulation parameters are zero.

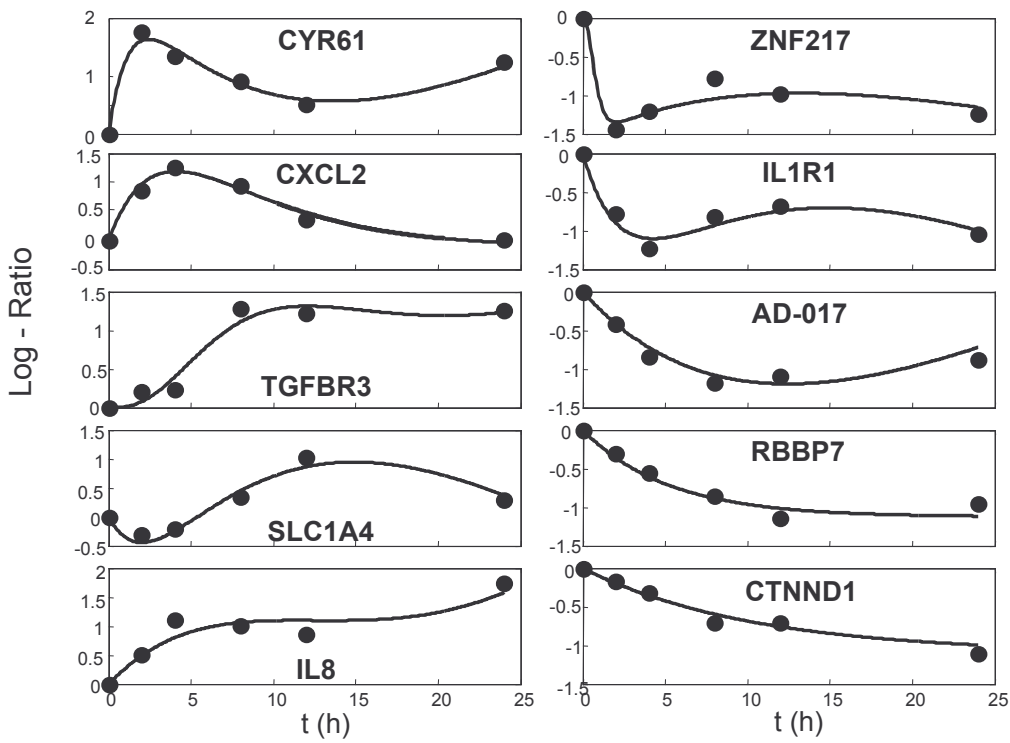


Fig. 3 Gene expression profiles of representative genes of the clusters. Measured data ( $\bullet$ ) and kinetics (continuous lines) simulated by the network model shown in Figure 2 as well as formulated in eq.(1) with the parameters listed in Table II

Target $i$					Source				
$i$	Symbol	$b_i$	valid	$w_{i,i}$	$R_i$	$j$	Symbol	$w_{i,j}$	valid
						$D_i$ in eq. (1)			
1	<i>CYR61</i>	+ 8.149	*	- 0.855	1	8	<i>AD-017</i>	+ 1.310	**
2	<i>CXCL2</i>	+ 3.169	***	- 0.287	1	9	<i>RBBP7</i>	+ 0.724	-
3	<i>TGFBR3</i>	+ 0.299	***	- 0.536	3	7	<i>IL1R1</i>	- 0.146	-
4	<i>SLC1A4</i>	- 2.150	*	- 0.443	1	8	<i>AD-017</i>	- 0.821	*
5	<i>IL8</i>	+ 1.206	**		1	8	<i>AD-017</i>	+ 0.254	**
6	<i>ZNF217</i>	- 300.8	***	- 3.692	3	8	<i>AD-017</i>	- 22.05	**
7	<i>IL1R1</i>	- 3.039	-	- 0.413	1	8	<i>AD-017</i>	- 0.403	-
8	<i>AD-017</i>	- 0.955	***		1	10	<i>CTNND1</i>	- 0.0968	*
9	<i>RBBP7</i>	- 0.872	*	- 0.194	1	-	-	-	-
10	<i>CTNND1</i>	- 0.426	***	- 0.097	1	-	-	-	-

Tab. II: Parameters of the model equation (see (1)): The external stimulus response vector is given by  $b_i$ , the self-regulation parameters  $w_{i,i}$  and the dynamic order  $R_i$  of the target gene  $i$  as well as the interaction matrix component  $w_{i,j}$  that quantifies the influence of source gene  $j$  on the target gene  $i$ . The asteriks in the column ‘valid’ represent the model validation result (interactions confirmed for several standard deviations  $\sigma$  of randomized data:  $\sigma = 0.05$ : \*, 0.20: \*\*, and 0.30: \*\*\*)

## DISCUSSION

Hypotheses about the signaling pathways after LiCl stimulation of liver cells were generated by clustering and network reconstruction from gene expression data. The biological relevance of the reconstructed network has to be discussed critically. The majority (86 %) of differentially expressed genes were found to be at least initially down-regulated after LiCl stimulation (614 of 716 differentially expressed genes belong to the clusters 6 - 10). The gene expression profiles of 524 genes (84 %) are monotonically decreasing up to the end of observation, i.e. over 24 hours. Thus, a more general repressive influence of LiCl seems to superpose the specific influence on the Wnt/ $\beta$ -catenin pathway of interest. Nevertheless, 102 genes are at least initially up-regulated, the majority (87, i.e. 85%) of them have an increasing kinetics up to 12 hours or more. One of the genes characterized by a monotonically increasing expression profile is coding for Interleukin 8 (IL8), which is known to be activated by  $\beta$ -catenin [9]. Only 2 genes (*CYR61* - ‘cysteine-rich, angiogenic inducer, 61’ and *CTGF* - ‘connective tissue growth factor’) have an early temporal expression maximum already 2 hours after stimulation. It has been reported that *CYR61* can activate the Wnt/ $\beta$ -catenin pathway by interaction with GSK3 $\beta$  [10]. The latter is inactivated by LiCl. Thus, the increase in the expression of *CYR61* and *IL8* corresponds with the stimulation of the Wnt/ $\beta$ -catenin pathway. The studied gene expression data alone are insufficient to reveal the causality and underlying mechanisms of the interaction between LiCl, *CYR61*, *IL8*, the Wnt/ $\beta$ -catenin pathway and the Glutamine synthetase (GS), the marker enzyme of liver zonation and heterogeneity. By bootstrapping with a disturbance that simulates the measurement error the direct responses of LiCl on 8 of the 10 clusters were confirmed. With  $\sigma = 0.30$  only the LiCl induced inactivations of *AD-017*, *ZNF217* and *CTNND1* and the activations of *CXCL2* and *TGFBR3* were identified as robust. According to the network model the cluster represented by AD-017 seems to play a central role. The activation role of this cluster on *CYR61* and the inactivation of the cluster represented by *ZNF217* by AD-017 were validated using randomized data with the standard deviation  $\sigma = 0.2$ . However, the direct stimulation of *CYR61* by LiCl could not be validated by the bootstrap analysis with the standard deviation  $\sigma = 0.2$ . Some of the data visualized by the reconstructed network might describe side effects of the LiCl stimulation, not involved in the heterogeneity of hepatocyte gene expression. The crucial point of the data-driven reverse engineering approach used is the selection of the biological correct cluster and the corresponding representative gene. In addition, some genes may be considered as network nodes as well. Thus, it was tested if a correlation of the 716 differentially expressed genes and the Wnt/ $\beta$ -catenin pathway are reported in literature. After reducing the cut-off for the identification of differentially expressed genes from the commonly accepted fold 2 as used in the present study to 1.62 (justified by estimation of the measurement error and biological variance) a larger number of differentially expressed genes was found including such

genes which are known to be involved in the activation of the Wnt/ $\beta$ -catenin pathway. This data- and knowledge-driven approach to signaling pathway reconstruction will be presented in a forthcoming paper [3].

## ACKNOWLEDGEMENT

This work was supported by the German Federal Ministry for Education and Research BMBF within the Program 'Systems of Life – Systems Biology' (FKZ 0313079B and FKZ 0313081).

## REFERENCES

- [1] Gebhardt R (1992) Metabolic zonation of the liver: regulation and implications for liver function. *Pharmacol Ther* 53: 275-354
- [2] Loeppen S, Schneider D, Gaunitz F, Gebhardt R, Kurek R, Buchmann A, Schwarz M (2002). Overexpression of glutamine synthetase is associated with beta-catenin-mutations in mouse liver tumors during promotion of hepatocarcinogenesis by phenobarbital. *Cancer Res.* 62: 5685-8.
- [3] Zellmer S, Schmidt-Heck W, Gaunitz F, Guthke R, Gebhardt R (2005). Dynamic network reconstruction from gene expression data describing the effect of LiCl stimulation on hepatocytes. *J. Integrative Bioinformatics*, accepted.
- [4] Kitano H (2002): Computational systems biology. *Nature* 420: 206 – 210.
- [5] Kitano H (2005): International alliances for quantitative modeling in systems biology. *Molecular Systems Biology*: Epub: March 29, 2005.
- [6] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004): Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5: R80, 2004.
- [7] Bolstad B (2004): affyPLM: Methods for fitting probe level models to Affy data. <http://www.bioconductor.org/repository/devel/vignette/affyPLM.pdf>
- [8] 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.
- [9] Levy L, Neuveut C, Renard CA, Charneau P, Branchereau S, Gauthier F, Van Nhieu JT, Cherqui D, Petit-Bertron AF, Mathieu D, Buendia MA (2002): Transcriptional activation of interleukin-8 by  $\beta$ -Catenin-Tcf4. *J Biol Chem*, 277: 42386–42393.
- [10] Latinkic BV, Mercurio S, Bennett B, Hirst EM, Xu Q, Lau LF, Mohun TJ, Smith JC (2003) : Xenopus Cyr61 regulates gastrulation movements and modulates Wnt signalling. *Development*, 130: 2429-2441.