

# **Integrative Modeling of Transcriptional Regulation in Response to Autoimmune Disease Therapies**

**Dissertation**  
zur Erlangung des akademischen Grades  
**doctor rerum naturalium**  
**(Dr. rer. nat.)**

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät  
der Friedrich-Schiller-Universität Jena

von  
**Diplom-Bioinformatiker**  
**Michael Hecker**  
geboren am 6. April 1982 in Erfurt

Die vorgelegte Arbeit, finanziert durch das Bundesministerium für Bildung und Forschung (Grant 0313692D), wurde am Leibniz-Institut für Naturstoff-Forschung und Infektionsbiologie e.V. - Hans-Knöll-Institut (HKI) unter der Leitung von PD Dr. Reinhard Guthke (Abteilung Systembiologie / Bioinformatik) im Zeitraum November 2006 bis Januar 2010 angefertigt.

# Table of contents

<b>Abbreviations.....</b>	<b>III</b>
<b>1. Introduction.....</b>	<b>1</b>
1.1. Gene regulatory network modeling.....	1
1.2. Autoimmune diseases.....	4
1.3. Objectives and experimental approach.....	8
<b>2. Overview of manuscripts.....</b>	<b>11</b>
<b>3. Manuscript I.....</b>	<b>16</b>
Hecker M, Lambeck S, Töpfer S, van Someren E, Guthke R.	
Gene regulatory network inference: data integration in dynamic models - a review.	
<i>Biosystems</i> 2009, 96(1):86-103.	
<b>4. Manuscript II.....</b>	<b>35</b>
Hecker M, Goertsches RH, Engelmann R, Thiesen HJ, Guthke R.	
Integrative modeling of transcriptional regulation in response to antirheumatic therapy.	
<i>BMC Bioinformatics</i> 2009, 10:262.	
<b>5. Manuscript III.....</b>	<b>54</b>
Goertsches RH, Hecker M, Koczan D, Serrano-Fernández P, Möller S,	
Thiesen HJ, Zettl UK.	
Long-term genome-wide blood RNA expression profiles yield novel molecular	
response candidates for IFN-beta-1b treatment in relapsing remitting MS.	
<i>Pharmacogenomics</i> 2010, 11(2):147-161.	
<b>6. Manuscript IV.....</b>	<b>70</b>
Hecker M, Goertsches RH, Fatum C, Koczan D, Thiesen HJ, Guthke R, Zettl UK.	
Network analysis of transcriptional regulation in response to intramuscular	
interferon-beta-1a multiple sclerosis treatment.	
<i>Pharmacogenomics J.</i> submitted January 25, 2010.	
<b>7. Discussion.....</b>	<b>104</b>

7.1. Discussion of main results.....	104
7.2. Discussion of methods.....	107
7.2.1. Experimental approach.....	107
7.2.2. Microarray data preprocessing.....	108
7.2.3. Integrative network inference.....	109
7.2.4. Evaluation of inference performance.....	111
7.3. Open issues and outlook.....	113
7.3.1. Prediction of clinical responses.....	113
7.3.2. Further development of TILAR.....	114
7.4. Concluding remarks.....	116
8. Summary.....	118
9. Zusammenfassung.....	120
References.....	122
Appendix.....	127
Danksagung.....	141
Ehrenwörtliche Erklärung.....	142
Tabellarischer Lebenslauf.....	143

## Abbreviations

ACPA	anti-citrullinated protein/peptide antibody
ACR	American College of Rheumatology
CDF	chip definition file
CSF	cerebrospinal fluid
DAS	disease activity score
DMARD	disease-modifying antirheumatic drug
DREAM	dialogue on reverse-engineering assessment and methods
EDSS	expanded disability status scale
GO	Gene Ontology
GRN	gene regulatory network
HLA	human leukocyte antigen
IFN- $\beta$	interferon-beta
IRF	IFN regulatory factor
LARS	least angle regression
Lasso	least absolute shrinkage and selection operator
MAID	MA-plot-based signal intensity-dependent fold-change criterion
MHC	major histocompatibility complex
MRI	magnetic resonance imaging
MS	multiple sclerosis
OLS	ordinary least squares
PBMC	peripheral blood mononuclear cells
PPI	protein-protein interaction
RA	rheumatoid arthritis
RF	rheumatoid factor
RNAP	RNA polymerase
ROC	recall-precision curve
RPC	receiver operating characteristic
TF	transcription factor
TFBS	transcription factor binding site
TILAR	TFBS-integrating least angle regression
TNF- $\alpha$	tumor necrosis factor-alpha

# 1. Introduction

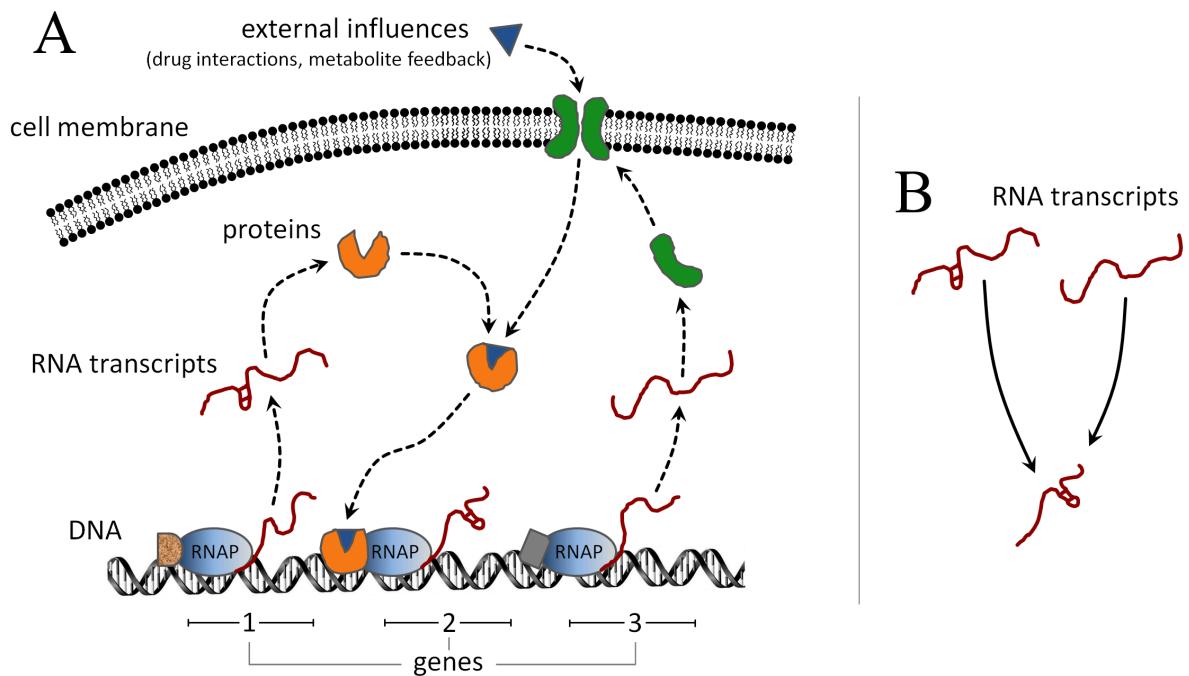
At the heart of multicellular life are the complex interactions between genes, proteins and metabolites. These interactions give rise to the function and behavior of biological systems. To study and understand such systems as a whole abstractions are needed such as the concept of networks, in which molecules are represented as nodes and interactions or causal influences are represented by edges. The reconstruction of biomolecular networks from experimental data and subsequent network analysis is a challenging and active field of research. The major focus of the present dissertation is on the inference of gene regulatory networks (GRNs).

## 1.1. Gene regulatory network modeling

Gene expression is mainly regulated at the level of DNA transcription by proteins called transcription factors (TFs). These TFs specifically bind short DNA sequence motifs at the regulatory region of their target genes. In doing so, they control the recruitment of RNA polymerase, which reads the DNA and transcribes it into RNA. However, gene regulation is a far more complex multi-layered process. Any step of gene expression may be modulated, from the RNA synthesis to the post-translational modification of proteins. Many genes are (directly oder indirectly) involved in these gene regulatory mechanisms, and therefore it is reasonable to regard genes as nodes in a network of mutual regulatory interactions.

The introduction of DNA microarrays in the mid-1990s offered the possibility to simultaneously measure the levels of thousands of RNA transcripts in a single sample of cells or tissues. Since then, researchers utilized the growing amount of large-scale gene expression data as input for algorithms to infer, or "reverse-engineer", the regulatory interaction structure of genes [1-3]. When inferring models of transcriptional regulation solely from gene expression data, one typically seeks for influences between RNA transcripts. In this case, the expression levels of each gene are explained by the expression levels of other genes. By construction, such GRN models do not generally describe physical interactions as transcripts rather exert their regulatory effects indirectly through the action of proteins, metabolites and effects on the cell environment (figure 1). Therefore, these models can be difficult to interpret in terms of real physical interactions, and the implicit description of hidden regulatory factors may limit the reliability of the inference results.

To overcome these issues and support the network reconstruction it is necessary to integrate additional biological information. Diverse types of data (e.g. protein-protein and protein-



**Figure 1.** Gene regulation is carried out by interactions of RNA molecules, proteins and metabolites. **(A)** Simplified illustration of an example GRN where gene "3" encodes a membrane-bound metabolite transporter protein (green shape). The metabolite (blue triangle) that is imported by this protein binds a TF (orange shape). The activated TF binds the DNA and together with RNA polymerase (RNAP) initiates the transcription of gene "2". Hence, the expression of gene "2" is influenced by the other two gene transcripts (red lines). **(B)** A graph model of the network in **(A)**. Because the model is inferred from measurements of RNA transcripts only, it implicitly captures the regulatory mechanisms at the protein and metabolite level that are not measured. The inferred GRN is a projection of the true network where transcripts influence the level of each other, even though they do not physically interact.

DNA interaction data) and prior knowledge (e.g. from experts, scientific literature and biological databases) can be analyzed in combination with gene expression data. This actually means that known or putative network links are preferred to be in the model. Furthermore, biological plausible assumptions about the network topology (e.g. structural sparseness) should be considered by including specific modeling constraints. Evidently, these advancements allow for GRN models that more accurately describe (physical) interactions between genes. The integration of heterogeneous data and use of prior knowledge is a current focus in the field of GRN inference in particular, and in computational biology in general.

Various modeling approaches have been proposed to reconstruct GRNs from experimental data. They rely on different mathematical concepts and learning strategies, and distinct degrees of abstraction. The most commonly used modeling formalisms are Boolean networks, Bayesian networks, association networks and systems of equations. Their main ideas as well as their pros and cons are described in more detail in manuscript I. This review addresses two major aspects: dynamic network models, i.e. models that have a time-component, and GRN inference methods that employ different types of information. Here it is sufficient to say that Bayesian networks and systems of linear equations (linear models) have been most studied for an integrative modeling. Linear models constitute the basic modeling framework used in this work (manuscript II and IV). They are therefore described briefly in the following.

Linear models assume that the regulatory relationship between genes is approximately linear. In principle, one can distinguish a static and a dynamic type of model:

- static model:  $x_i = \sum_{j=1, j \neq i}^N w_{ij} x_j + b_i + \varepsilon_i$
- dynamic model:  $x_i[t+1] - x_i[t] = \sum_{j=1}^N w_{ij} x_j[t] + b_i + \varepsilon_i[t]$

where  $x_i$  is the expression of gene  $i$  (at time  $t$ ),  $N$  is the number of genes in the network,  $w_{ij}$  denotes an edge weight, i.e. the regulatory effect of gene  $j$  on gene  $i$ , and  $b_i$  represents a (possible) external influence on gene  $i$ . As each experimental observation will contain some error, a "disturbance term"  $\varepsilon_i$  is introduced adding noise to the linear influence function. The dynamic form (a system of linear difference equations) is preferred if appropriate time-course gene expression data are available, and if not only the network structure but also the dynamics are considered important, e.g. for simulation purposes. In comparison to static models, dynamic models also allow for self-interactions that express the sum of RNA degradation and self-regulation.

When inferring a static or dynamic linear GRN model as outlined above one needs to estimate the model parameters  $w_{ij}$  and  $b_i$  from the data. However, despite the simplicity of these models, their inference is always challenging. This is because of the combinatorial nature of the problem, the incomplete knowledge of the molecules involved in specific conditions, and because experimental data are typically limited and noisy. Amount and quality of the gene expression data have a strong impact on the reliability of a GRN reconstruction. Moreover, in case of an integrative modeling, the model accuracy depends on

the quality of the additional biological information to be utilized and the capability of the inference algorithm to incorporate such (possibly uncertain) information (see also section "data requirements" in manuscript I).

Once derived, the structure of the network can be analyzed in more detail. Characteristic for GRNs is the presence of hubs (key genes regulating multiple genes) and specific network motifs (frequently recurring interaction patterns) as well as a modular network structure (where sets of genes are highly interconnected and share a given property) [4]. Beyond that, the study of gene interactions could provide insights into control principles such as redundancy and feedback, and uncover intertwined gene regulatory cascades. Dynamic models allow to evaluate the network's dynamical stability. They can also be applied to predict the time behavior of a system for different parameter settings. As a result, an inferred GRN model often provides a bunch of new hypotheses and generates assumptions for further research activities. Therefore, network modeling can be useful in many applications. In computational medicine, an important issue is to unveil and examine the architecture of GRNs under normal and pathological conditions. Here, the inference of regulatory interactions between genes can play a pivotal role in understanding the mechanisms, diagnosis and treatment of complex diseases such as human autoimmune diseases [5].

## 1.2. Autoimmune diseases

Autoimmune diseases are disorders characterized by an inappropriate immune response against constituents normally present in the human body. More exactly, the immune system produces T cells and antibodies that attack cells, tissues and organs of the body as if they were foreign, thereby leading to inflammation and damage. The causes of autoimmune conditions are largely unknown, but it appears that there is an inherited genetic predisposition in many cases [6]. Autoimmune diseases are typically multifactorial polygenic diseases and affect approximately 3% of the world population [7]. To better understand such complex disorders it is crucial to unravel the structure and dynamics of the molecular networks that play a role in the aberrant immune response. Network analyses may not only support the investigation of autoimmune diseases, but also the optimization of their treatment. Two common diseases with an autoimmune basis are rheumatoid arthritis (RA) and multiple sclerosis (MS) (table 1).

RA is a chronic inflammatory disorder primarily afflicting the synovial joints. Blood-derived cells migrate into the joints and together with activated synovial cells produce cytokines and degradative enzymes that progressively lead to the destruction of cartilage and bone. The

**Table 1.** Comparative overview of RA and MS disease attributes.

		Multiple sclerosis	Rheumatoid arthritis
Characteristics	Target tissue	Central nervous system (brain and spinal cord)	Synovial joints and other tissues / organs
	Attributes	Chronic, inflammatory, T cell mediated, autoimmune	Chronic, inflammatory, systemic, autoimmune
	Course	Relapsing-remitting, progressive	Progressive, often fluctuating disease activity
	Pathophysiology	Immune cells cross the blood-brain barrier and induce neurodegeneration (axonal and myelin loss)	Inflammation of synovial membrane leads to cartilage destruction and ankylosis of joints (polyarthritis); immune dysregulation affects the whole body
	Etiology	Unknown	Unknown
	Risk factors	Genetic: MHC locus (e.g. HLA-DRB1*1501 allele <sup>1</sup> ) Environmental: infectious agents, smoking, stress	Genetic: MHC locus (e.g. HLA-DRB1*0401 allele <sup>1</sup> ) Environmental: infectious agents, smoking
Outcomes	Symptoms	Cognitive impairment, physical disability, muscle weakness, deficits in sensation and of movement, visual and speech problems	Joint swelling and pain, significant disability, loss of mobility, morning stiffness, low-grade fever, malaise, fatigue, loss of appetite
	Life expectancy	Normal	Reduction by 5 - 10 years
Epidemiology	Prevalence (European population)	0.06 - 0.2 %	0.5 - 1.0 %
	Concordance: first-degree relatives	2 - 5 %	2 - 3 %
	Concordance: monozygotic twins	20 - 35 %	15 - 30 %
	Female : male ratio	1.6 - 2.0 : 1	2 - 3 : 1
	Age at onset (years)	20 - 40	35 - 50
	Geographic distribution	Less common in people living near the equator or in countries with lower socioeconomic level	Consistent throughout the world with some exceptions
Clinical management	Diagnosis	McDonald criteria, CSF analysis	ACR criteria, ACPA tests
	Prognosis	Depends on disease subtype, sex, age and initial symptoms	Depends on initial symptoms (early joint damage), RF and ACPA status, sex and age
	Measures of disease severity	EDSS, number of relapses, MRI	DAS, radiology
	Treatment	Immunosuppressive, immuno-modulatory (e.g. IFN-β administration, B cell depletion)	Immunosuppressive, immuno-modulatory (e.g. TNF-α inhibition, B cell depletion)
	Treatment goals	Control symptoms, prevent progression	Control symptoms, prevent progression, relief of pain

Abbreviations: ACPA = anti-citrullinated protein/peptide antibody, ACR = American College of Rheumatology, CSF = cerebrospinal fluid, DAS = disease activity score, EDSS = expanded disability status scale, MHC = major histocompatibility complex, MRI = magnetic resonance imaging, RF = rheumatoid factor.

<sup>1</sup> The human leukocyte antigen (HLA) complex is involved in antigen presentation, a process crucial to the initiation of an adaptive immune response.

disease is systemic and over the years more and more joints are affected. Significant disability and a reduced quality of life are result of chronic disease activity [8].

MS is a central nervous system disease. It is the most common progressive and disabling neurologic disease of young adults. In people with MS, lesions accumulate in the brain and

spinal cord, particularly in the white matter, and damage the myelin covering of nerve fibres. Inflammation and the loss of myelin cause disruption to nerve transmission and thus affect many functions of the body [9,10].

In the pathogenesis of both, RA and MS, genetic susceptibility, environmental exposure and immune dysregulation play significant roles. However, it is less clear how the chronic inflammation is set up and maintained exactly. It is assumed that the onset of these diseases in predisposed individuals may also reflect random processes during immune cell development such as immunoglobulin or T cell receptor gene recombination and mutation [11]. Common genes are likely to be involved in both diseases, as the chromosomal regions that contain susceptibility genes coincide [12]. Nevertheless, MS patients do not have a higher risk to develop RA and vice versa. In fact, a reduced comorbidity of RA and MS has been observed [13]. Both diseases follow an unpredictable course with variable severity. Our incomplete understanding of autoreactive inflammatory processes, the clinical heterogeneity of the diseases, and their complex pathogeneses are among the factors that render a specific treatment very challenging.

It is presently not possible to cure RA, MS or any other autoimmune disorder. Nevertheless, the course of RA and MS can be relatively well controlled in many patients by disease-modifying therapies. Current medications aim to alleviate symptoms, minimize organ and tissue damage, prevent functional loss, and slow the advance of the disease, generally by decreasing the immune response. Existing MS therapies are primarily designed to limit lesion formation and brain atrophy, reduce the rate and severity of relapses (acute periods of worsening) and slow progression to disability [14]. The goal of RA management is to relieve pain and to prevent further joint damage [15].

The use of immunomodulatory "biologic" agents has significantly improved the treatment of these diseases. Biologics, in contrast to drugs that are chemically synthesized, are derived from living sources. They are often complex proteins targeting the disease in a more specific manner than traditional therapies. Several biologic drugs for RA and MS try to intervene in the immune system by altering the levels of cytokines. Cytokine proteins play a critical role as mediators of immune regulation. The tumor necrosis factor-alpha (TNF- $\alpha$ ) cytokine is a master regulator of inflammation, and has been widely implicated in the pathophysiology of both diseases [16,17]. Clinical trials have shown that blocking the action of TNF- $\alpha$  reduces disease activity in RA patients [18,19], but not in MS patients [20]. Biologic TNF- $\alpha$  antagonists are usually prescribed to patients with RA when other disease-modifying antirheumatic drugs (DMARDs) have failed to work, or have produced intolerable side

effects. For controlling the exacerbations in relapsing-remitting MS, interferon-beta (IFN- $\beta$ ) administration is currently the most established treatment and several studies have confirmed its clinical benefit [21,22]. IFN- $\beta$  is, like TNF- $\alpha$ , a natural human pleiotropic cytokine. Its antiproliferative activities are believed to be responsible for the beneficial effect of IFN- $\beta$ -based drugs. In addition, IFN- $\beta$  improves the integrity of the blood-brain barrier, which generally breaks down in MS patients [23].

However, these (and other) biologics are only modestly effective and fail to control disease progression in a subset of patients. Approximately one-third of MS patients suffers from a higher or identical annual relapse rate while on IFN- $\beta$  treatment than before (so-called non-responders) [24]. Similarly, only 60% percent of RA patients receiving a TNF- $\alpha$  blocking agent for at least 6 months achieve a 20% improvement in the ACR criteria [25]. Adverse effects and the occurrence of drug-neutralizing antibodies in some patients are further problems related to these therapies and also demonstrate the patients' heterogeneity in therapeutic response.

To improve the clinical success of RA and MS therapies, more individualized treatment approaches are needed. For this purpose, it is essential to early identify the patients that benefit most from particular therapies. This would protect the patients against unnecessary drug exposure, loss of time and side effects, improve compliance and probably reduce health care expenditures. However, currently no clinically established laboratory (molecular) markers are available that allow to predict either favorable or detrimental therapeutic outcomes with sufficient reliability. Therefore, it is important to better understand the pharmacogenomic and -dynamic effects of these drugs.

Although IFN- $\beta$  and anti-TNF- $\alpha$  agents have been marketed for more than a decade, we are still learning how they work. While the intracellular signaling pathways triggered by TNF- $\alpha$  and IFN- $\beta$  are known in great detail [15,26], their effects on gene expression in the individual are relatively poorly understood. Eventually, they influence many immune system processes - directly or indirectly as a consequence of the activity of induced proteins. Ongoing research continues to explore the precise molecular mechanisms that are instrumental for the drugs' therapeutic efficacy and the side effects associated with them. Network models can disclose useful information about the various complex processes involved. Therefore, GRN inference techniques are suited to investigate the therapeutic effects on transcriptional regulation, and may help to explain why some patients do not achieve adequate clinical responses.

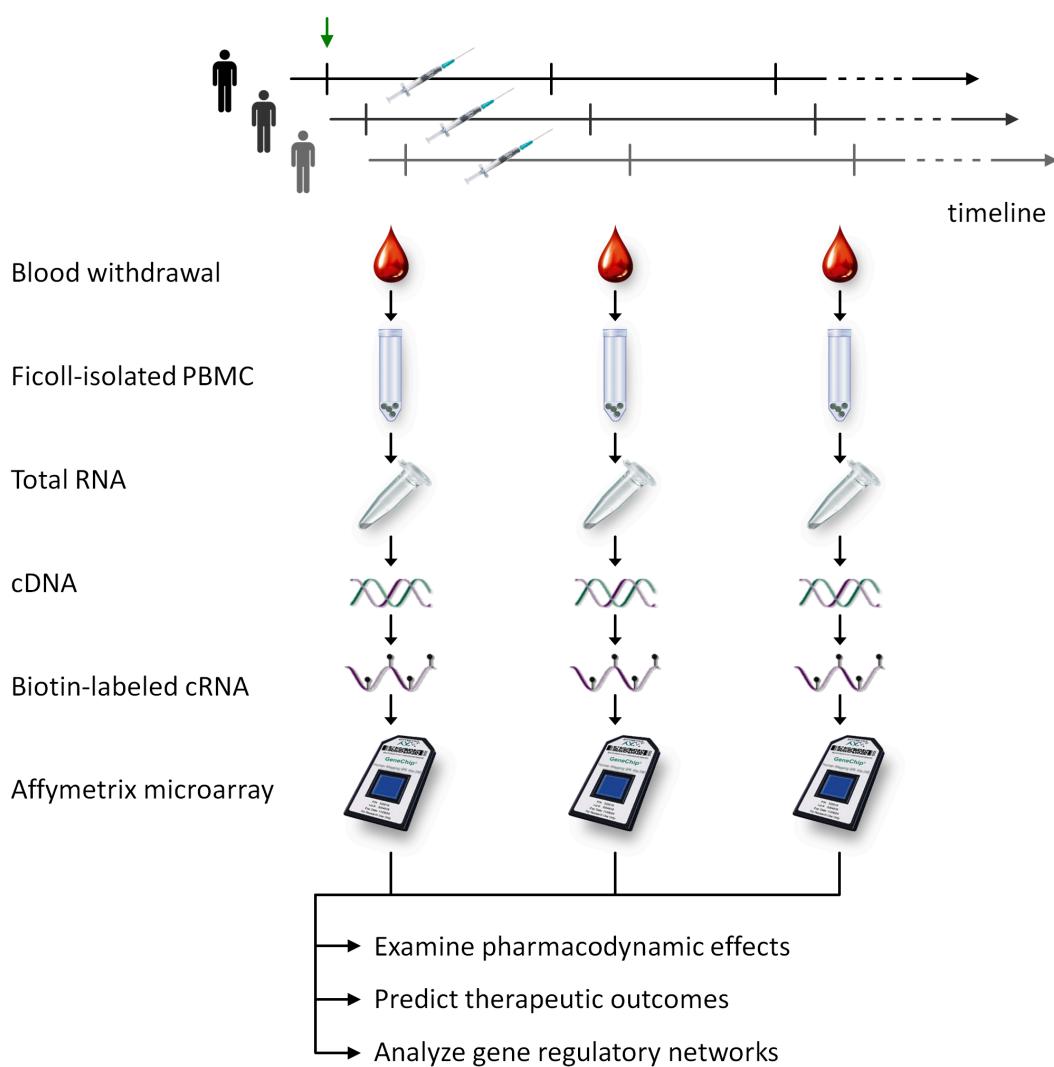
### 1.3. Objectives and experimental approach

Etanercept (Enbrel, Wyeth) is an example of a protein-based drug directed against TNF- $\alpha$ . It is a soluble TNF- $\alpha$  receptor fusion protein used to treat moderate to severe RA and other inflammatory disorders. Betaferon (IFN- $\beta$ -1b subcutaneous, Bayer Schering) and Avonex (IFN- $\beta$ -1a intramuscular, Biogen Idec) are two similar but different IFN- $\beta$  medications for patients with relapsing-remitting form of MS. The aim of this work was to characterize the transcriptional effects induced by these immunomodulatory therapies. Gene network analyses were applied for elucidating the interactions between genes responsive to the different drugs (manuscript II-IV).

For each therapy, genome-wide expression profiles of a group of patients were measured using Affymetrix DNA microarrays. The general experimental workflow used to obtain the data is depicted in figure 2. As shown, gene expression levels were determined in peripheral blood mononuclear cells (PBMC) immediately before therapy initiation and at different time points during treatment. The PBMC population (lymphocytes and monocytes) is the main cellular source of inflammatory cytokines and thus plays a major role in the immune system. Blood here serves as a surrogate tissue, i.e. as a sensor of autoimmune diseases which actually affect other, less accessible tissues.

As a first step to the analysis, genes significantly up- or down-regulated after start of drug administration had to be identified. Next, by screening the regulatory regions of the genes for known TF binding sites (TFBS), one can reveal TFs that are putatively responsible for the transcriptional changes observed in the data. The aim was then to reverse-engineer regulatory interactions between the genes based on gene expression data, TFBS information and/or text-mining knowledge (that is knowledge gathered from biological literature by computational means). As a result, novel hypotheses about the drugs' mode of action were expected from the reconstructed GRN models.

In the Betaferon study (manuscript III), molecular interactions obtained by text-mining were used to build network structures. For unraveling the regulatory effects between genes up- or down-regulated in response to Etanercept and Avonex therapy, an integrative GRN inference algorithm, called TILAR, was developed and applied to the data (manuscript II and IV). The algorithm realizes a new modeling constraint to derive a model from the expression data while directly incorporating TFBS information. As it employs a linear modeling approach, the method can draw on well established statistical techniques to efficiently fit the model parameters.



**Figure 2.** Study design to investigate transcriptional regulation in response to three different therapies. Peripheral venous blood was taken from a group of patients immediately before first (green arrow) and selected subsequent drug injections. The blood samples were then processed in different steps to perform an Affymetrix microarray analysis. First, whole RNA was extracted from isolated PBMC. Second, through reverse transcription, the RNA was converted into double-stranded complementary DNA (cDNA), which in turn served as a template to generate biotinylated cRNA. Next, this labeled cRNA was randomly fragmented and added to the microarrays. After hybridization of the cRNA fragments to complementary oligonucleotide probes on the chip, the arrays were washed, stained with fluorescent molecules that stick to Biotin and scanned with a laser. The detected fluorescence intensities disclose what genes were expressed and at what approximate level. Finally, the obtained expression profiles can be analyzed bioinformatically. Given clinical data of the patients one can seek for genes whose expression before or early in therapy is predictive for long-term benefit of the respective treatments. The focus of the presented works was to characterize blood gene expression changes during therapy and to apply GRN inference to improve the understanding about the drugs' mechanisms of action.

Analysis of network topology was then performed to uncover sub-networks of co-regulated and functionally related genes as well as network motifs, e.g. feedback and feedforward loops. Such network properties provide useful insights into the complex regulatory processes underlying the drugs' therapeutic effects. Finally, the inferred GRN structures were correlated to clinical data (e.g. treatment response and drug-related side effects) to identify gene sets that could serve an individual prognosis of therapeutic outcomes. A further task was to evaluate the performance of the TILAR modeling algorithm to investigate how far integrating different types of biological data can improve the inference results.

## 2. Overview of manuscripts

### MANUSCRIPT I (PUBLISHED)

#### **Gene regulatory network inference: data integration in dynamic models - a review**

Hecker M, Lambeck S, Töpfer S, van Someren E, Guthke R

*Biosystems* 2009, 96(1):86-103.

#### **Summary**

This is a review on the reconstruction of GRNs from experimental data through computational methods. It focuses on dynamic network modeling approaches and the integration of prior biological knowledge and heterogeneous types of data. First, some background on the nature of large-scale gene expression experiments together with short descriptions of methods for analyzing such data is given. Basic modeling frameworks are then specified and the mathematical principles which both define and limit them. After addressing their pros and cons, algorithms for GRN inference under those models are described. General difficulties of GRN reconstruction are discussed as well as ways to counter them. Key modeling properties, like low average connectivity, and integrative modeling strategies that aid in understanding and inference of GRNs are presented. Moreover, the review covers the validation of inferred models and the evaluation of network reconstruction methods, and outlines future perspectives in GRN modeling.

#### **Authors' contributions**

All authors compiled the literature, drafted parts of the manuscript and prepared the tables and figures. MH was particularly involved in writing chapter 2 and 6. Chapter 5 was mainly written by ST, and RG wrote the summary in chapter 8. RG and MH completed and improved the whole manuscript.

## MANUSCRIPT II (PUBLISHED)

### **Integrative modeling of transcriptional regulation in response to antirheumatic therapy**

Hecker M, Goertsches RH, Engelmann R, Thiesen HJ, Guthke R.

*BMC Bioinformatics* 2009, 10:262.

#### **Summary**

The purpose of this study was to characterize the transcriptional effects induced by Etanercept therapy in patients with RA. The analysis is based on an Affymetrix DNA microarray dataset providing genome-wide PBMC expression profiles of 19 RA patients within the first week of treatment. Significant transcriptional changes were observed for 83 genes. The so-called TILAR algorithm was introduced for the first time in this work and applied on the data to model the regulatory interactions between these genes. First of all, overrepresented predicted TFBS were identified in the genes' regulatory regions. TILAR then allowed to integrate the gene expression data and TFBS information to infer a model of the underlying GRN. A hybrid of the least angle regression and the ordinary least squares regression was used to find the model structure and estimate the parameters. The reconstructed GRN exhibits a scale-free and self-regulating organization, and indicates the pleiotropic immunological role of the therapeutic target TNF- $\alpha$ . A benchmarking analysis demonstrates that TILAR is able to reconstruct GRNs more reliably than other established methods, in particular if additional prior knowledge on the regulators of TF activity is available and incorporated during inference (adaptive TILAR).

#### **Authors' contributions**

RG and HJT directed the study. MH conceived and implemented the algorithms, carried out the analyses on the data, wrote the paper and prepared all tables and figures. RG, RHG and RE assisted in interpretation of the results and corrected and improved the paper.

### MANUSCRIPT III (PUBLISHED)

#### **Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN-beta-1b treatment in relapsing remitting MS**

Goertsches RH, Hecker M, Koczan D, Serrano-Fernández P, Möller S, Thiesen HJ, Zettl UK  
*Pharmacogenomics* 2010, 11(2):147-161.

#### **Summary**

In this work, long-term transcriptional profiles were obtained for 25 patients with relapsing-remitting MS. Using Affymetrix microarrays, PBMC expression levels were measured for each patient before (baseline), as well as two days, one month, one year and two years after start of Betaferon therapy. The data were analyzed with the aim to investigate the pharmacodynamic effects of sustained drug administration. The strongest changes to the transcriptome were observed at one month into treatment: A total of 175 genes were significantly up- or down-regulated in comparison to baseline. Nineteen genes were consistently found modulated over the two years observation period which suggests that the medication continually influences the patients' immune system. The lists of filtered genes were finally examined for overrepresented gene functional annotations, and literature text-mining was used to explore molecular interactions between the genes. Two major gene networks were identified: the first consists of several known IFN- $\beta$ -induced genes, whereas the second mainly contains down-regulated genes that to date have not been associated with IFN- $\beta$  activity. The study thus provides novel marker genes for biological response to Betaferon and information on the molecular networks it affects.

#### **Authors' contributions**

UKZ and HJT inspired and directed the work. The lab experiments were performed by DK. RHG was responsible for data analysis and interpretation, writing the paper as well as preparing tables and figures. MH actively supported the data evaluation, and participated in discussion of the results and preparation of the manuscript. UKZ was involved in patient care and, together with S-FP and SM, contributed to writing the paper.

## MANUSCRIPT IV (SUBMITTED)

### **Network analysis of transcriptional regulation in response to intramuscular interferon-beta-1a multiple sclerosis treatment**

Hecker M, Goertsches RH, Fatum C, Koczan D, Thiesen HJ, Guthke R, Zettl UK

*Pharmacogenomics J.* submitted January 25, 2010.

#### **Summary**

The aim of this study was to investigate the transcriptional effects induced by Avonex therapy in patients with relapsing-remitting MS. For 24 MS patients, Affymetrix DNA microarrays were employed to determine the gene expression levels of PBMC within the first four weeks of IFN- $\beta$  administration. Overall, 121 genes were found significantly up- or down-regulated during therapy, including known IFN- $\beta$ -modulated genes. Analysis of the regulatory regions of these genes revealed 11 overrepresented TFBS. As in the RA study (manuscript II), the new TILAR algorithm was then applied for deriving a GRN model from the gene expression data and TFBS predictions. The inferred network shows a scale-free topology and specifies a number of feedback loops. An NF- $\kappa$ B-centered sub-network of genes was found higher expressed in patients with IFN- $\beta$ -related side effects. The inferred GRN thus provides novel insights into functional mechanisms of Avonex therapy in MS, and exposes molecular differences between the patients. A benchmarking analysis confirmed that the integrative modeling strategy realized by TILAR performs much better than algorithms using gene expression data or TFBS information alone.

#### **Authors' contributions**

RG and UKZ directed the study. CF and UKZ were responsible for patient care and clinical documentation. DK participated in performing the microarray and real-time PCR experiments. MH conducted the analysis and interpretation of the data, wrote the paper and prepared all tables and figures. RG, RHG and HJT assisted in interpretation of the results and contributed to the writing of the paper.

### Overview of authors' percentage contributions

<b>Titel</b>	<b>Journal</b>	<b>Autoren</b>	<b>Arbeitsanteil</b>
Gene regulatory network inference: data integration in dynamic models - a review.	Biosystems 2009, 96(1):86-103.	Hecker M Lambeck S Töpfer S van Someren E Guthke R	35% 20% 10% 10% 25%
Integrative modeling of transcriptional regulation in response to antirheumatic therapy.	BMC Bioinformatics 2009, 10:262.	Hecker M Goertsches RH Engelmann R Thiesen HJ Guthke R	88% 4% 2% 2% 4%
Long-term genome wide blood RNA expression profiles yield novel molecular response candidates for interferon beta-1b treatment in relapsing remitting multiple sclerosis.	Pharmacogenomics accepted October 29, 2009.	Goertsches RH Hecker M Koczan D Serrano-Fernández P Möller S Thiesen HJ Zettl UK	50% 24% 14% 1% 1% 1% 9%
Network analysis of transcriptional regulation in response to intramuscular interferon-beta-1a multiple sclerosis treatment.	Pharmacogenomics J. submitted January 25, 2010.	Hecker M Goertsches RH Fatum C Koczan D Thiesen HJ Guthke R Zettl UK	72% 5% 1% 12% 2% 4% 4%

### **3. Manuscript I**

---

#### **Gene regulatory network inference: data integration in dynamic models - a review**

Michael Hecker, Sandro Lambeck, Susanne Töpfer, Eugene van Someren,  
and Reinhard Guthke

*Biosystems* 2009, 96(1):86-103.

---



Contents lists available at ScienceDirect

**BioSystems**journal homepage: [www.elsevier.com/locate/biosystems](http://www.elsevier.com/locate/biosystems)

## Gene regulatory network inference: Data integration in dynamic models—A review

Michael Hecker<sup>a</sup>, Sandro Lambeck<sup>a</sup>, Susanne Toepfer<sup>b</sup>, Eugene van Someren<sup>c</sup>, Reinhard Guthke<sup>a,\*</sup>

<sup>a</sup> Leibniz Institute for Natural Product Research and Infection Biology - Hans Knoell Institute, Beutenbergstr. 11a, D-07745 Jena, Germany

<sup>b</sup> BioControl Jena GmbH, Widenbruchstr. 15, D-07745 Jena, Germany

<sup>c</sup> Centre for Molecular and Biomolecular Informatics (CMBI) and Department of Applied Biology, Nijmegen Centre for Molecular Life Sciences, Radboud Universiteit Nijmegen, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

### ARTICLE INFO

#### Article history:

Received 12 March 2008

Received in revised form 5 November 2008

Accepted 9 December 2008

#### Keywords:

Systems biology

Reverse engineering

Biological modelling

Knowledge integration

### ABSTRACT

Systems biology aims to develop mathematical models of biological systems by integrating experimental and theoretical techniques. During the last decade, many systems biological approaches that base on genome-wide data have been developed to unravel the complexity of gene regulation. This review deals with the reconstruction of gene regulatory networks (GRNs) from experimental data through computational methods. Standard GRN inference methods primarily use gene expression data derived from microarrays. However, the incorporation of additional information from heterogeneous data sources, e.g. genome sequence and protein–DNA interaction data, clearly supports the network inference process. This review focuses on promising modelling approaches that use such diverse types of molecular biological information. In particular, approaches are discussed that enable the modelling of the dynamics of gene regulatory systems. The review provides an overview of common modelling schemes and learning algorithms and outlines current challenges in GRN modelling.

© 2008 Elsevier Ireland Ltd. All rights reserved.

### 1. Introduction

In ‘systems biology’, one aims to model the physiology of living systems as a whole rather than as a collection of single biological entities. Such an approach has the practical benefit of offering insight into how to control or optimise parts of the system while taking into account the effect it has on the whole system. Therefore, taking a ‘systems-wide’ view may lead to alternative solutions in application areas such as biotechnology and medicine. The ability to take a systems-wide approach is only possible due to recent developments in high-throughput technologies that enable scientists to carry out global analyses on the DNA and RNA level and large-scale analyses on the protein and metabolite level. To gain a better understanding of the observed complex global behaviour and the underlying biological processes, it is necessary to model the interactions between a large number of components that make up such a biological system. To be able to learn respective large-scale models, the use of novel computational methods that can make an integrative analysis of such different sources of data is essential and challenging at the same time.

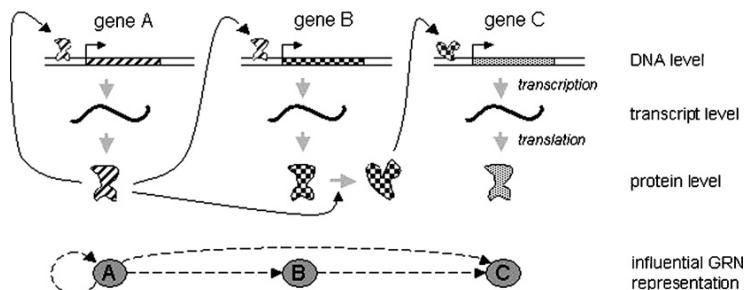
Uncovering the dynamic and intertwined nature of gene regulation is a focal point in systems biology. The activity of a gene’s func-

tional product is influenced not only by transcription factors (TFs) and co-factors that influence transcription, but also by the degradation of proteins and transcripts as well as the post-translational modification of proteins. A gene regulatory network (GRN) aims to capture the dependencies between these molecular entities and is often modelled as a network composed of nodes (representing genes, proteins and/or metabolites) and edges (representing molecular interactions such as protein–DNA and protein–protein interactions or rather indirect relationships between genes). Many GRN inference approaches solely consider transcript levels and aim to identify regulatory influences between RNA transcripts. Such approaches employ an ‘influential’ GRN, i.e. a GRN where the nodes consist of genes and edges represent direct as well as indirect relationships between genes (Fig. 1). This approximation leads to ‘influence’ network models that are intended to implicitly capture regulatory events at the proteomic and metabolomic level which sometimes makes them difficult to interpret in physical terms. The modelling (reconstruction) of a GRN based on experimental data is also called reverse engineering or network inference. Reverse engineering GRNs is a challenging task as the problem itself is of a combinatorial nature (find the right combination of regulators) and available data are often few and inaccurate.

Therefore, it is beneficial to integrate system-wide genomic, transcriptomic, proteomic and metabolomic measurements as well as prior biological knowledge (e.g. from the scientific literature) into a single modelling process. Using computational support to

\* Corresponding author. Tel.: +49 3641 532 1083; fax: +49 3641 532 0803.

E-mail address: [reinhard.guthke@hki-jena.de](mailto:reinhard.guthke@hki-jena.de) (R. Guthke).



**Fig. 1.** Schematic view of a simple gene regulatory network. Gene A regulates its own expression and those of gene B. Thereby, gene A might exert its regulatory influence directly (if it encodes a TF) or indirectly (if it controls the activity of another TF possibly via a signalling cascade). When reconstructing the GRN, one often aims to infer an ‘influence’ network model as shown at the bottom.

adequately manage, structure and employ heterogeneous types of information in order to obtain a more detailed insight into biological network mechanisms represents a major challenge in GRN inference today.

Outstanding review articles covering the field of data-driven inference of GRNs are from De Jong (2002), van Someren et al. (2002a), Gardner and Faith (2005), Filkov (2005), Van Riel (2006), Bansal et al. (2007), Goutsias and Lee (2007), Cho et al. (2007) as well as Markowetz and Spang (2007). Well-structured overviews of the general idea behind GRN inference and diverse common mathematical modelling schemes can be found in De Jong (2002) and Filkov (2005). van Someren et al. (2002a) arranged reverse engineering techniques according to the characteristics of their underlying model and learning strategies; moreover, the pros and cons of distinct approaches are discussed. Gardner and Faith (2005) clearly outlined between two general reverse engineering strategies: (1) physical models that describe real physical interactions such as TF–DNA interactions and (2) influence models that allow any type of influence to be modelled, but do not necessarily provide a physical explanation of an effect. Markowetz and Spang (2007) focused on probabilistic models, such as Bayesian networks.

In this review we want to emphasize two major aspects: dynamic network models, i.e. approaches that aim to capture the complex phenomena of biological systems by modelling the time-course behaviour of gene expression, and integration of prior biological knowledge and heterogeneous sources of data. We chose the following text structure according to the main steps taken during the modelling of GRNs (Fig. 2): first, experimental aspects and biological databases relevant to the study of GRNs are addressed, and main issues of data-driven modelling discussed. Next, Section 3 provides a survey of typical GRN modelling architectures. Section 4

deals with data- and knowledge-driven feature selection and mapping methods which aim at reducing the number of variables in the model to lower model complexity. Fundamental learning strategies for inferring GRNs are described in Section 5. In Section 6 we focus on inference methods that employ other types of data in addition to gene expression measurements. Section 7 addresses the validation of inferred mathematical models and the assessment of network inference methods. Section 8 draws conclusions and outlines perspectives for future research on GRN inference.

## 2. Biological Data

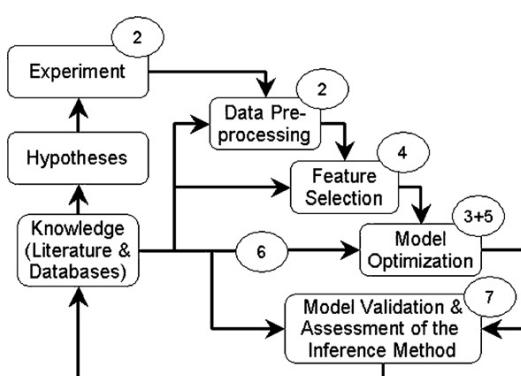
The reconstruction of GRNs is largely promoted by advances in high-throughput technologies, which enable to measure the global response of a biological system to specific interventions. For instance, large-scale gene expression monitoring using DNA microarrays is a popular technique for measuring the abundance of mRNAs. However, by integrating different types of ‘omics’ data (e.g. genomic, transcriptomic and proteomic data) the quality of network reconstruction could be drastically improved. In the following we outline the main characteristics of diverse ‘omics’ data, itemize distinct types of molecular interactions and briefly refer to relevant databases as well as measurement techniques.

### 2.1. Omics Data and Related Technologies

Genome sequence data are supportive to the reconstruction of GRNs since transcription is regarded as the main control mechanism of gene expression. The analysis of sequence data includes the investigation of TF binding sites (TFBS). Thereby, the aim is to detect potential links between sequence motifs and tissue-specific gene expression. In the past, a vast amount of *in silico* approaches has been developed to identify TF binding sequence elements (Wasserman and Sandelin, 2004). However, as computational approaches provide only a simplified representation of DNA-binding events, usually a large number of potential binding sites (candidates) are predicted, of which many are not functional (false positives). More detailed data on TFs and their binding sites are accessible via databases such as JASPAR and TRANSFAC.

Experimentally, the ChIP-on-chip technique (chromatin immunoprecipitation combined with microarray technology) allows to characterise protein–DNA interactions at high-throughput. By identifying the regions of a genome that are bound by a particular TF *in vivo*, potential gene regulatory effects can be derived (Ren et al., 2000). As more and more ChIP-on-chip data (also called location data) are generated, the large number of *in silico* predicted TFBSs gets more and more paralleled by a large number of experimentally observed TFBSs.

The amounts of transcripts, proteins and metabolites available at a specific point in time reflect the current state (of activity) of



**Fig. 2.** The systems biology cycle. In this cycle, knowledge leads to new hypotheses, which leads to new experiments, which leads to new models, which leads to new knowledge, etc. Numbers refer to the corresponding sections.

**Table 1**

Categories of interaction databases presented in Pathguide as of 10/2008. In the column '#Resources' the number of databases belonging to each category is shown.

Category	Content	#Resources	Examples
Protein–protein interactions	Mainly pairwise interactions between proteins	105	DIP, BIND, STRING, HPRD
Metabolic pathways	Biochemical reactions in metabolic pathways	60	KEGG, Reactome, ENZYME
Signaling pathways	Molecular interactions and chemical modifications in regulatory pathways	50	STKE, Reactome, TRANSPATH
Transcription factors/Gene regulatory networks	Transcription factors and the genes they regulate	42	JASPAR, TRANSFAC, RegulonDB
Pathway diagrams	Hyperlinked pathway images	30	KEGG, HPRD, SPAD
Protein–compound interactions	Interactions between proteins and compounds	24	ResNet, CLIBE
Protein sequence focused	Diverse pathway information in relation with sequence data	16	REBASE
Genetic interaction networks	Genetic interactions, such as epistasis	6	BIND, BioGRID

the biological system. *Transcriptome* data measured by genome-wide DNA microarrays are traditionally used for GRN modelling as RNA molecules are easily accessible in comparison to proteins and metabolites. In general, two types of DNA microarrays can be distinguished: single and two-channel microarrays. Thereby, the abundance of mRNAs is typically quantified on the basis of short DNA oligonucleotides and cDNA molecules, respectively (Kawasaki, 2006). A huge amount of gene expression microarray data is publicly available via repositories such as ArrayExpress and Gene Expression Omnibus. However, one has to be aware that DNA microarray data are characterised by a high degree of variability (noise). One way to overcome this problem is to apply real-time quantitative PCR assays (RTQ-PCR) to get more precise measures of transcript levels for a selected set of genes.

Similar to the transcriptome, the term *proteome* describes the ensemble of proteins produced in a cell or organism. Protein levels are decisively influenced by the amount of mRNA transcripts. Remarkably, the total number of human proteins is much higher than the number of protein-encoding human genes, because alternative mRNA splicing and post-translational processing increase the proteome diversity. Moreover, proteins often form complexes with other proteins or RNA molecules to achieve specific function and activity. The structural variety of proteins and their functional interactions cause a high degree of complexity and therefore large-scale proteomic studies are usually difficult (Panedy and Mann, 2000).

Proteins (enzymes) can catalyse enzymatic reactions and thus are also the basis of all metabolic events. Metabolites also modulate GRNs, however, similarly as within proteomics, technical difficulties hamper a global analysis of the *metabolome* (Goodacre et al., 2004). Therefore, connecting metabolic and gene regulatory networks is out of the scope of this review and remains a challenge for the future. However, it should be noted that the area of systems biological modelling originates from the modelling of metabolic networks (Heinrich and Schuster, 1996).

A complementary approach to the systematic measurement of molecular and cellular states is the characterization of molecular interactions. The complex network of intermolecular interactions that wires together the vast amount of genes, proteins and small molecules is also called the *interactome*. Here, high-throughput methods enable researchers to systematically screen for protein–protein and protein–DNA interactions (e.g. ChIP-on-chip for the latter as mentioned above). Interactome information can also be found in many different databases containing known and predicted interactions. Some of these databases provide detailed information on regulatory proteins and their associated regulated genes (e.g. RegulonDB for *Escherichia coli*), others contain known metabolic networks (e.g. KEGG), still others catalogue protein–protein interaction (PPI) information (e.g. DIP). More than

260 web-accessible biological pathway and network databases are linked in the meta-database Pathguide (Table 1). Note that the information in these databases is by far not complete.

However, besides the wealth of information stored in biological databases, a large amount of information is found in the scientific literature. Therefore, text mining tools have been developed to automatically extract interrelations between genes and proteins from literature with sufficient reliability (e.g. PathwayStudio). Clearly, such text mining methods also provide useful information for GRN modelling.

A further type of data relevant to study genes and their regulatory interactions are gene functional annotations. Several projects such as the Gene Ontology (GO) provide a controlled vocabulary to describe gene and gene product attributes. The functional annotations in the GO database (GO terms) are organized in a hierarchical way defining subsets of genes that share common biological functions. This type of information alleviates the functional interpretation of genes participating in a GRN.

Clearly, this section does not provide a complete and detailed description of the diverse types of biological data that are available. However, it illustrates the potential benefit as well as the challenge of utilizing such diverse and complementary types of biological information to reliably infer GRNs.

## 2.2. Experimental Design

As a gene expression experiment is often the basis for a GRN reconstruction, some aspects concerning the design of such experiments will be covered here. Specifying the experimental design is an important issue in the investigation of GRNs, since the choice of a modelling approach and its learning strategy is often related to the type and amount of data generated. At least two aspects are crucial in this context: perturbation (i.e. the choice of intervention or experimental condition) and observation (sampling, measurement) of the biological system.

### 2.2.1. Perturbation

In general, systems identification is based on the analysis of input–output signal data that describe the system's response to perturbations (Ljung, 1999). Similarly, in order to understand a dynamic biological system, i.e. its behaviour and functioning, we need to perturb it systematically. Perturbation experiments can be designed in different ways depending on the available techniques and the system of interest, and include manipulations of environmental factors as well as interventions on the genetic, transcriptomic, proteomic or metabolic level.

Environmental perturbations comprise, e.g. heat shock, chemical stresses or compound-treatments up to the administration of therapeutic agents in medical care. In comparison, genetic

perturbations, e.g. gene deletion (knock-out) and over-expression studies, may exclusively affect those genes in the network, which lay downstream of the perturbed gene and are therefore a valuable method to specifically detect regulatory dependencies. However, genetic perturbation experiments are not easy to establish in most organisms. The realisation of such studies is therefore restricted to *in vitro* experiments or to lower organisms such as the eukaryotic model organism *Saccharomyces cerevisiae* (yeast).

In addition, experiments including perturbations on the transcriptome level can be performed and used for GRN inference (Markowetz et al., 2005; Rice et al., 2005). One possibility is to use a natural mechanism called RNA interference, which is an RNA-guided regulation of gene expression (so-called knock-down experiment) (Fire et al., 1998; Mello and Conte, 2004).

Having techniques available that can directly intervene on the molecular level, researchers are in the position to selectively affect gene expression. The ability to systematically influence the expression of genes in a network as well as to subsequently measure altered gene expression levels also allows for alternating arrangements of experimental design and model construction. Ideker et al. (2000), for example, proposed a network inference approach in which genes with the most uncertain connections in the network model are perturbed in order to incrementally determine a Boolean network (see Section 3.2) using only a few experiments. The underlying concept of iterated and systematic perturbations was also used by Tegnér et al. (2003). Although a promising strategy, the applicability of this approach to real data remains to be proven, since both authors used artificial data in their work.

### 2.2.2. Sampling

Effects of interventions can be observed by static (steady state) or time-course measurements, where the latter reflects the dynamic behaviour of the system over time. Therefore, the choice between a static and a dynamic GRN model largely depends on the experimental setup. The setup, in turn, should depend on the type of knowledge one aims to achieve, i.e. the importance of capturing the effect that changes in initial conditions have on the final states (static) versus capturing the sequence of intermediate processes that leads to the final state (dynamic).

While generating *static* data (at well-defined experimental conditions), the observation is accomplished at the presumed steady state of the biological system. For instance, samples taken from knock-out organisms are supposed to provide gene expression levels at steady-state in the absence of a specific gene product. Network inference based on data derived from knock-out experiments is very efficient (Bansal et al., 2007). However, to infer all interactions from such data, each node in the network has to be perturbed separately. Moreover, the steady-state design may miss dynamic events that are critical for reliably inferring the structure of a GRN.

On the other hand, *time-series* experiments (when samples are taken in a series of time-points after perturbation) might reveal dynamics, but the data may contain redundant information leading to inefficient use of experimental resources. It should be noted that an appropriate design of time-series experiments is difficult on its own. For instance, one has to find a compromise between observation duration and the interval between two subsequent measurements, as the number of time-points also determines the amount of experimental efforts. Note that the number and allocation of time-points for sampling affects the performance of the GRN inference as was studied using synthetic data (Yeung et al., 2002; Geier et al., 2007).

### 2.3. Data Requirements

While generating experimental data, researches have to face a trade-off. On the one hand, they aim to minimise experimental

efforts and costs, hence try to minimise the number of experiments. On the other hand, a reliable GRN reconstruction cannot be done without a considerable quantity of accurate data.

The general opinion is that the amount of data required for GRN modelling (e.g. DNA microarrays) increases approximately logarithmically with the number of network nodes (e.g. genes) (Akutsu et al., 1999; Yeung et al., 2002; Filkov, 2005). However, it is difficult to specify the experimental data requirements more precisely as many further factors influence the network inference performance.

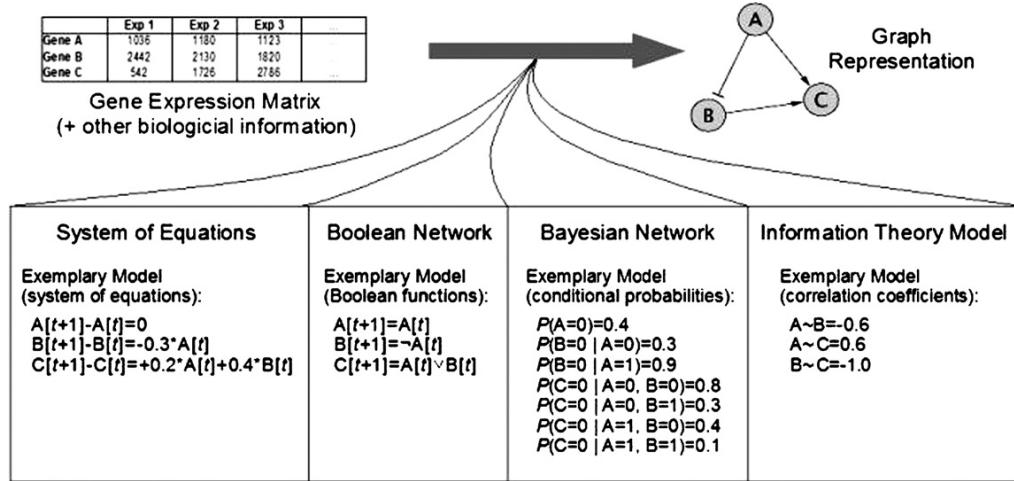
One, the quality of an inferred model depends on the quality of the given data. Large variations in the biological outcome, high measurement noise and inappropriate experimental designs might lead to less informative data and thus hamper a reliable GRN reconstruction. Two, the aim of the modelling can range from estimating gene regulatory interactions with high confidence up to reconstructing even highly speculative regulatory dependencies. Precise estimates of parameters are not always needed to understand certain qualitative features of a GRN. Three, different modelling formalisms exhibit different data requirements. More complex models consist of many model parameters and therefore their learning is more data demanding. Four, different network inference algorithms infer gene regulatory effects from a given amount of data with different efficacy. Searching for the best model parameter setting is typically computationally intractable even for simple models. Hence, heuristics have to be applied (see Section 5), which may perform suboptimally. Moreover, the applied inference technique might exploit modelling constraints such as sparseness of the inferred network (see below and Section 6.1) or assess the accuracy of the edges in the network by internal validation (see Section 7.2) to increase the inference reliability. Five, the inference strategy might use external prior knowledge from databases and literature (see Section 6). In this case, the necessary amount of experimental data depends on the amount, type and quality of such additional information and the capability of the inference algorithm to adequately integrate this information during modelling.

To summarize, there is a tight relationship between model complexity, the amount/type of data required for inference and the quality of the results. Due to this, the inference of more accurate (i.e. complex, dynamic, large-scale) GRN models is impeded. The main problem is that a more accurate modelling makes the correct model much harder to find, because the size of the search space increases exponentially with the number of unknown model parameters (the so-called problem of dimensionality). In consequence, the modeller has to counter the dimensionality problem in network inference, for example by:

- (i) increasing the amount of data by increasing the number of measurements  $M$ ;
- (ii) reducing the number of network nodes  $N$ ;
- (iii) restricting the number of model parameters, e.g. by use of simple models and network connectivity constraints;
- (iv) integrating specific prior knowledge about the network structure.

(i), the number of measurements  $M$  can be increased by additional experiments or by merging own gene expression data with complementary data from external repositories, e.g. as in Faith et al. (2007). Alternatively, D'haeseleer et al. (1999) proposed for time-series data to simply interpolate additional time-points between the actual measured time-points. This is justified by the fact that gene expression levels change rather smoothly over time. However, it was shown that interpolation did little to solve the dimensionality problem (Wessels et al., 2001).

(ii), the number of network nodes  $N$  can be reduced by focusing on features (genes, proteins, ...) of special interest employing



**Fig. 3.** Exemplary overview of the four main GRN modelling architectures. Here, the aim is to infer the regulatory interactions of three genes (the GRN graph on the top-right) based on the expression data of these three genes for a handful of experiments (the gene expression matrix on the top-left). For modelling we may utilize (pre-processed) gene expression data as well as other available biological information. The four modelling architectures reflect the same GRN in different ways. Each of the model architectures shown here is illustrated by a single typical example of a possible realization of its model formalism category. Note that there exist a lot more approaches for each of the model architectures than what is reflected by the given examples.

methods for feature selection and/or feature mapping (see Section 4).

(iii), by using less complex network models and biologically motivated modelling constraints, the dimensionality of the model search space can be reduced. The most widely used modelling constraint is the sparseness constraint, which minimises the number of edges in the network thereby reducing the number of (unknown) model parameters (see Section 6.1).

(iv), the integration of different types of biological data may augment the modelling and thus facilitates inference results of higher quality (see Section 6.2).

#### 2.4. Data Pre-processing

Data pre-processing is a critical step in GRN reconstruction as it affects the performance of the inference algorithms and thus the inference results, i.e. the generated hypotheses. Methods for data pre-processing have to be applied specifically to different data types and experimental designs.

The analysis of large-scale data is a challenging task, not so much because the amount of data is large, but because large-scale measurement technologies possess high inherent variability. The two sources of this variability are systematic errors (bias) and stochastic effects (noise). Systematic effects affect all measurements in a similar manner and thus can be nearly eliminated by data normalisation (Quackenbush, 2002). Stochastic effects cannot be corrected by pre-processing, but can be quantified, in particular by the application of repeated measurements (replicates).

Depending on the modelling approach further data manipulations may be necessary. Several network inference methods require a very specific pre-processing of the data. For instance, interpolation of time-series data is a frequently applied method. Furthermore, many learning algorithms for the inference of differential equation systems (see Section 3.3) require the estimation of time derivatives for each measurement point of the time-series, which can also be done by interpolation (Chen et al., 1999; Yeung et al., 2002). Besides, some network formalisms require discrete gene expression values. For instance, to infer Boolean networks, the measured expression levels have to be converted into binary numbers. Note that such a data discretization is often non-trivial and has to be done with adequate care.

### 3. Network Model Architecture

Before inferring a GRN, the appropriate type of network model architecture has to be chosen. The model architecture is a parameterised mathematical function that describes the general behaviour of a target component based on the activity of regulatory components. Once the model architecture has been defined, the network structure (i.e. the interactions between the components) and the model parameters (e.g. type/strengths of these interactions) need to be learned from the data (see Section 5). Over the last years, a number of different model architectures for reverse engineering GRNs from gene expression data have been proposed. They cover varying degrees of simplification and reflect distinct assumptions of the underlying molecular mechanisms (Fig. 3).

In general, the network nodes represent compounds of interest, e.g. genes, proteins or even modules (sets of compounds). As described by van Someren et al. (2002a), model architectures can be distinguished by (1) the representation of the activity level of the network components. The concentration or activity of a compound can be represented by Boolean ('on', 'off') or other logic values (e.g. 'present', 'absent', 'marginal'), discrete (e.g. cluster labels), fuzzy (e.g. 'low', 'medium', 'high') or continuous (real) values. Furthermore, network model architectures can be distinguished by (2) the type of model (stochastic or deterministic, static or dynamic) and (3) the type of relationships between the variables (directed or undirected; linear or non-linear function or relation table). Although many undirected network representations exist, the focus of this review is on directed networks.

#### 3.1. Information Theory Models

One of the simplest network architectures is the correlation network (Stuart et al., 2003), which can be represented by an undirected graph with edges that are weighted by correlation coefficients. Thereby, two genes are predicted to interact if the correlation coefficient of their expression levels is above some set threshold. The higher the threshold is set, the sparser is the inferred GRN.

Besides correlation coefficients, also Euclidean distances and information theoretic scores, such as the mutual information, were applied to detect gene regulatory dependencies (Steuer et

al., 2002). The network inference algorithms RELNET (RELevance NETworks; Butte and Kohane, 2000), ARACNE (Algorithm for the Reverse engineering of Accurate Cellular NEtworks; Margolin et al., 2006; Basso et al., 2005) and CLR (Context Likelihood of Relatedness; Faith et al., 2007) apply network schemes in which edges are weighted by statistic scores derived from the mutual information. Rao et al. (2007) proposed an asymmetric version of the mutual information measure to obtain directed networks. Likewise, graphical Gaussian models (GGMs) using partial correlations to detect conditionally dependent genes also allow to distinguish direct from indirect associations (Opgen-Rhein and Strimmer, 2007).

Simplicity and low computational costs are the major advantages of information theory models. Because of their low data requirements, they are suitable to infer even large-scale networks. Thus, they can be used to study global properties of large-scale regulatory systems. In comparison to other formalisms, a drawback of such models is that they do not take into account that multiple genes can participate in the regulation. A further disadvantage is that they are static.

### 3.2. Boolean Networks

Boolean networks are discrete dynamical networks. They were first proposed by Kauffman (1969) and since then have been intensively investigated for modelling gene regulation (Thomas, 1973; Bornholdt, 2008). They use binary variables  $x_i \in \{0, 1\}$  that define the state of a gene  $i$  represented by a network node as ‘off’ or ‘on’ (inactive or active). Hence, before inferring a Boolean network, continuous gene expression signals have to be transformed to binary data. The discretization can be performed, for instance, by clustering and thresholding using support vector regression (Martin et al., 2007). Boolean networks can be represented as a directed graph, where the edges are represented by Boolean functions made up of simple Boolean operations, e.g. AND ( $\wedge$ ), OR ( $\vee$ ), NOT ( $\neg$ ). The challenge of reverse engineering a Boolean network is to find a Boolean function for each gene in the network such that the observed (discretised) data are explained by the model. Various algorithms exist for the inference of Boolean networks, e.g. REVEAL (REVerse Engineering ALgorithm; Liang et al., 1998). REVEAL was later extended to allow for multiple discrete states as well as to let the current state depend not only on the prior state but also on a window of previous states.

Boolean networks are limited by definition as gene expression cannot be described adequately by only two states. Nevertheless, Boolean networks are easy to interpret and as they are dynamic, they can be used to simulate gene regulatory events. In naïve Boolean network models there are no kinetic constants and other continuous variables.

### 3.3. Differential and Difference Equations

Differential equations describe gene expression changes as a function of the expression of other genes and environmental factors. Thus, they are adequate to model the dynamic behaviour of GRNs in a more quantitative manner. Their flexibility allows to describe even complex relations among components. A modelling of the gene expression dynamics may apply ordinary differential equations (ODEs):

$$\frac{dx}{dt} = f(x, p, u, t) \quad (1)$$

where  $x(t) = (x_1(t), \dots, x_n(t))$  is the gene expression vector of the genes  $1, \dots, n$  at time  $t$ ,  $f$  is the function that describes the rate of change of the state variables  $x_i$  in dependence on the model parameter set  $p$ , and the externally given perturbation signals  $u$ .

Here, network inference means the identification of function  $f$  and parameters  $p$  from measured signals  $x, u$  and  $t$ .

In general, without constraints, there are multiple solutions, i.e. the ODE system is not uniquely identifiable from data at hand. Thus, the identification of model structure and model parameters requires specifications of the function  $f$  and constraints representing prior knowledge, simplifications or approximations. For instance, the function  $f$  can be linear or non-linear. Evidently, regulatory processes are characterised by complex non-linear dynamics. However, many GRN inference approaches based on differential equations consider linear models or are limited to very specific types of non-linear functions (Voit, 2000; De Jong, 2002; see Section 3.3.2).

There are further, more complex variants of differential equation models, such as stochastic differential equations that are thought to take into account the stochasticity of gene expression, which might occur especially when the number of TF molecules is low (Kaern et al., 2005; Climescu-Haulica and Quirk, 2007).

#### 3.3.1. Linear Differential Equations

A linear model:

$$\frac{dx_i}{dt} = \sum_{j=1}^N w_{i,j} \cdot x_j + b_i \cdot u, \quad i = 1, \dots, N \quad (2)$$

can be applied to describe the gene expression kinetics  $x_i(t)$  of  $N$  genes by  $N \times (N + 1)$  parameters for (a) the  $N^2$  components  $w_{i,j}$  of the interaction matrix  $W$  and (b)  $N$  parameters  $b_i$  quantifying, for example, the impact of the perturbation  $u$  on gene expression. In general, the simplification obtained by linearization is still not sufficient to identify large-scale GRNs from gene expression data unequivocally. Several approaches have been proposed to cope with this problem, e.g. methods for inferring sparse interaction matrices by reducing the number of non-zero weights  $w_{i,j}$  (see Section 5.2).

Differential equations can be approximated by difference equations (discrete-time models). Thereby, the linear differential Eq. (2) becomes the linear difference Eq. (3):

$$\frac{x_i[t + \Delta t] - x_i[t]}{\Delta t} = \sum_{j=1}^N w_{i,j} \cdot x_j[t] + b_i \cdot u, \quad i = 1, \dots, N \quad (3)$$

In this way one obtains a linear algebraic equation system that can be solved by well-established methods of linear algebra. Singular value decomposition (SVD) (Holter et al., 2001; Yeung et al., 2002) and regularised least squares regression are the most prominent ones that solve the linear equation system with the constraint of sparseness of the interaction matrix. For instance, the LASSO (Least Absolute Shrinkage and Selection Operator) provides a robust estimation of a network with limited connectivity and low model prediction error (van Someren et al., 2002b; see Section 5). Further inference algorithms based on linear difference equation models are NIR (Network Identification by multiple Regression; Gardner et al., 2003), MNI (Microarray Network Identification; di Bernardo et al., 2005) and TSNI (Time-Series Network Identification; Bansal et al., 2006). Under the steady-state assumption, NIR and MNI use series of steady-state RNA expression measurements, whereas TSNI uses time-series measurements to identify gene regulatory interactions (see also Bansal et al., 2007).

#### 3.3.2. Non-linear Differential Equations

Complex dynamic behaviours such as the emergence of multiple steady states (e.g. healthy or disease states) or stable oscillatory states (e.g. calcium oscillations and circadian rhythms) cannot be explained by simple linear systems. Instead, systems of cellular regulation are non-linear (Savageau, 1970, Heinrich and Schuster, 1996). The identification of non-linear models is not only limited

by mathematical difficulties and computational efforts for numerical ODE solution and parameter identification, but also mainly by the fact that the sample size  $M$  is usually too small for the reliable identification of non-linear interactions. Thus, the search space for non-linear model structure identification has to be stringently restricted. For that reason, inference of non-linear systems employ predefined functions that reflect available knowledge. [Sakamoto and Iba \(2001\)](#) used genetic programming to identify small-scale networks (up to three genes) by fitting polynomial functions  $f$  of differential Eq. (1). [Spieth et al. \(2006\)](#) applied different search strategies, such as evolutionary algorithms, for the inference of small-size networks (2, 5 and 10 genes). They studied different types of non-linear models: generalized linear network models ([Weaver et al., 1999](#)), S-systems ([Savageau, 1970](#); [Kimura et al., 2005](#)) and models composed of a linear interaction matrix and an additional non-linear term (called ‘H-systems’).

Exemplarily, S-systems model the gene expression rate by excitatory and inhibitory components:

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^N X_j^{g_{i,j}} - \beta_i \prod_{j=1}^N X_j^{h_{i,j}} \quad (4)$$

Here,  $\alpha_i$  and  $\beta_i$  are positive rate constants and  $g_{i,j}$  and  $h_{i,j}$  are kinetic exponents. Non-linear models such as S-systems consist of many parameters demanding a large number of experiments to fit them to the data ([Vilela et al., 2008](#); [Voit, 2008](#)). Therefore, the problem of data insufficiency still limits the practical relevance of non-linear models.

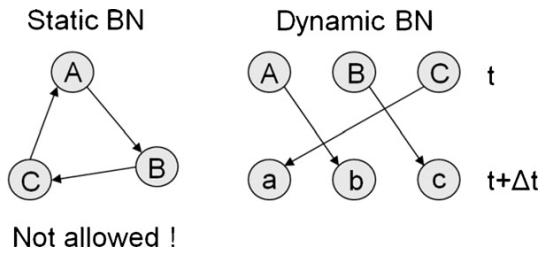
### 3.4. Bayesian Networks

Bayesian networks (BNs) reflect the stochastic nature of gene regulation and make use of the Bayes’ rule. Here, the assumption is that gene expression values can be described by random variables, which follow probability distributions. As they represent regulatory relations by probability, BNs are thought to model randomness and noise as inherent features of gene regulatory processes ([Friedman et al., 2000](#)). Most importantly, BNs provide a very flexible framework for combining different types of data and prior knowledge in the process of GRN inference to derive a suitable network structure ([Werhli and Husmeier, 2007](#); see also Section 6.2). Besides, BNs have a number of features that make them attractive candidates for GRN modelling, such as their ability to avoid over-fitting a model to training data and to handle incomplete noisy data as well as hidden variables (e.g. TF activities). Methods for learning BNs are covered in detail in [Heckerman \(1996\)](#) and [Needham et al. \(2007\)](#). In short, there are three essential parts for learning a BN:

- **Model selection.** Define a directed acyclic graph (DAG) as candidate graph of relationships.
- **Parameter fitting.** Given a graph and experimental data find the best conditional probabilities (CP) for each node.
- **Fitness rating.** Score each candidate model. The higher the score, the better the network model (the DAG and the learned CP distribution) fits to the data. The model with the highest score represents the GRN inference result.

Thereby, the critical step is ‘model selection’. The naïve approach is to simply enumerate all possible DAGs for the given number of nodes (so-called brute-force search). Unfortunately, the number of DAGs on  $N$  nodes, grows super-exponentially. Therefore, as for other model types, heuristics are needed to efficiently learn a BN (see Section 5).

BNs can be learned based on discrete (often Boolean) and continuous expression levels. Thereby, the underlying probabilistic model might be, e.g. a multinomial distribution or a Gaussian distribution.



**Fig. 4.** Difference between static BNs (left panel) and dynamic BNs (right panel). A feedback loop from gene A to gene B to gene C and back to gene A is not allowed in static BNs. However, this feedback loop can be represented in a dynamic BN.

BNs of continuous nodes are typically harder to infer from experimental data, because of their additional computational complexity. However, their inference does not require discretisation of the data. Moreover, static and time-series data can be used to reconstruct static and dynamic Bayesian networks (DBNs), respectively. As the former have the structure of a DAG, they cannot capture feedback loops. In contrast, DBNs separate input nodes from output nodes, i.e. each molecular entity is represented by a regulator node (representing the expression level at time  $t$ ) as well as by a target node (representing the expression level at time  $t + \Delta t$ ) ([Van Berlo et al., 2003](#); [Perrin et al., 2003](#)). This way, DBNs are able to describe regulatory feedback mechanisms, because a feedback loop will not create a cycle in the graph (Fig. 4).

BNs are widely used for GRN reconstruction (see also Section 6). As an example, [Rangel et al. \(2004\)](#) inferred a 39-gene linear state-space model – a subclass of DBNs – of T-cell activation from gene expression time-series data. Noteworthy, BANJO is a ready-to-use software application for BN and DBN inference ([Hartemink et al., 2001](#)).

### 3.5. Further Network Model Architectures

Not all GRN modelling techniques can be assigned to one of the four categories described above. To complete this section, three of these approaches are mentioned here exemplarily: [Segal et al. \(2003\)](#) identified regulatory modules in *S. cerevisiae* and used them for modelling the regulation program by regression trees. Thereby, each decision node in the tree corresponds to a regulating gene. The so-called Dynamic Regulatory Events Miner (DREM) algorithm introduced by [Ernst et al. \(2007\)](#) uses hidden Markov models for identification and annotation (by TF names) of so-called bifurcation points in gene expression profiles. As a third example, [Mordelet and Vert \(2008\)](#) decomposed the GRN inference into a large number of local binary classification problems, which focus on separating target genes from non-targets for each TF.

## 4. Feature Selection and Feature Mapping

To reliably identify the structure and parameters of a model, the model size/complexity must suit the experimental data at hand. In essence, both feature selection as well as feature mapping reduce the complexity of the model by selecting only relevant features for network reconstruction. While analysing gene expression data, genes that are non-responsive or not well measured in the data are typically removed during *feature selection*. With *feature mapping* molecular entities can be combined into functional entities that represent the common behaviour of its constituents or that reflect a particular biological function. Thus, a functional entity might be for instance a cluster of co-expressed genes or a group of proteins with the same function. Feature mapping is an excellent way

to remove redundant information. However, the modeller has to carefully choose which dimensionality reduction approach is appropriate to (a) obtain a sufficiently large network to investigate the biological phenomena under study while (b) still being able to obtain a reliable inference of the underlying network. Filtering differentially expressed genes and clustering co-expressed genes are widely applied techniques to reduce the number of model variables. Advanced feature selection/mapping approaches combine data- and knowledge-driven methods.

#### 4.1. Data-driven Feature Selection

Network reconstruction approaches often consider only genes that show significant changes in expression under the experimental conditions studied. For instance, Wang et al. (2006) narrowed down the list of relevant genes of *S. cerevisiae* to 140 genes based on 2-fold change up or down in at least 20% of the expression levels across all data sets. Guthke et al. (2005) selected 1336 cDNA features (out of 18,432 cDNAs representing 7619 unique genes) by requiring a 8-fold up- or downregulation after perturbation by infection. van Someren et al. (2006) studied 101 murine genes (out of 9596) that showed significant changes in expression with respect to the initial state under their experimental conditions. Martin et al. (2007) selected murine genes represented by 5085 probesets (out of 45,119 probesets representing ~34,000 unique genes) that exhibited differences in expression between control cells and IL-2-stimulated cells using the following inclusion criteria: (1) change call other than 'no change', (2) same trend of change call ('increase', 'decrease'), (3) 'present call' and 'signal intensity > 100' and (4) at least a 1.5-fold difference in expression between the two compared conditions. As a remark, the significance of expression change, often used for filtering candidate genes, can be assessed using *t*-statistics or its variations (Pan, 2002).

#### 4.2. Data-driven Feature Mapping

Another way to reduce the number of network components is the identification of clusters of co-expressed and/or co-regulated genes or proteins. Methods for cluster analysis have been widely applied to find functional groups under the assumption that genes which show similar expression patterns are co-regulated or part of the same regulatory pathway. Afterwards, cluster-representative genes or the mean expression level of all genes in a cluster might be used for GRN inference (D'haeseleer et al., 2000; Wahde and Hertz, 2000; Mjolsness et al., 2000; van Someren et al., 2000; Guthke et al., 2005, 2007; Bonneau et al., 2006).

Clustering does not guarantee that genes within a cluster share the same biological function. Nevertheless, a common subsequent analysis step is to annotate each cluster with a functional category that is representative for that cluster (Gibbons and Roth, 2002). From a statistical learning perspective, clustering methods can be subdivided into (a) combinatorial algorithms, (b) mixture modelling, and (c) mode seeking (Hastie et al., 2001). Hierarchical algorithms are still frequently employed, although they have been criticized (Morgan and Roy, 1995; Radke and Möller, 2004) and more reliable methods are available. For instance, *k*-means and fuzzy *c*-means (Granzow et al., 2001; Dougherty et al., 2002) were used in conjunction with GRN inference (Guthke et al., 2005). Apart from that, Mjolsness et al. (2000) applied an expectation-maximization algorithm for clustering by mixture modelling and used the mean time-courses of 'aggregated genes' for inferring a dynamic network model.

A complete description of clustering algorithms is beyond the scope of this review, and the reader is referred to the literature for more on this subject (Shannon et al., 2003).

#### 4.3. Knowledge-driven Feature Selection/Mapping

As feature selection/mapping is a crucial step in GRN inference, one might not only exploit the limited set of gene expression data but also employ alternative sources of biological information. One way to do this is to use knowledge about which genes code for transcription factors. For instance, one can start to select (known or putative) TFs, which are differentially expressed or just belong to a certain process of interest. Then, further genes can be additionally selected on the basis of their (known or putative) regulation by one or more of these TFs, e.g. as done by Bernard and Hartemink (2005). This selection can be based on protein–DNA binding data or based on results from searching for regulatory motifs in sequence information. However, a drawback of this feature selection approach is that the activity of TFs not necessarily correlates with their changes in transcript abundance.

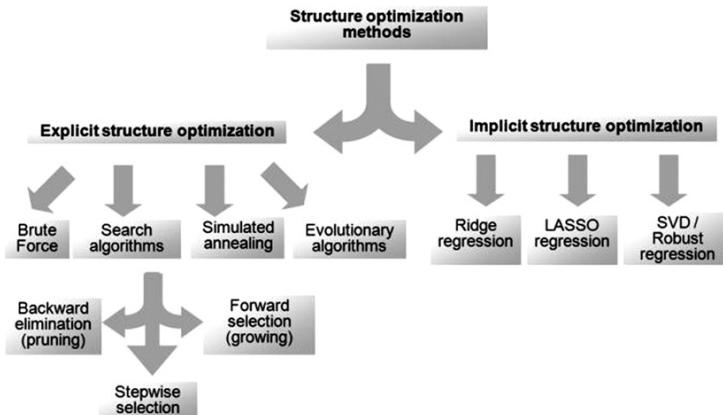
Alternatively, one can focus on modelling particular pathways or biological processes. Here, annotation databases (see Section 2.1; Table 1) provide functional classifications that can be used to directly select genes of a specific pathway, process or cellular component (see for an example: Hartemink et al., 2002). In analogy, one can select genes that are associated with the same biological context based on text mining (e.g. Tamada et al., 2003). A drawback of these solely knowledge-based approaches is that gene expression levels are not taken into account, and thus relevant features, which are not yet correctly annotated might be missed, while features that do not play a role under the particular conditions might be falsely included.

A more sophisticated way to reduce the number of features is to analyse the expression of specific groups of genes instead of individual genes. Using annotation terms in conjunction with expression levels allows to find functional modules, which play a key role in the particular system. Current methods that deduce a biological meaning, i.e. an association to functions and processes, from large-scale gene expression data, consist of two steps. At first, a group of genes is defined (e.g. by data-driven feature selection/mapping). Then, the enrichment of biologically relevant terms (derived from annotation databases) in these genes can be determined. For example using Gene Ontology one can test whether particular functions or processes are specifically related to the group of genes. A lot of freely available tools are based on this approach (Khatri and Draghici, 2005). These and other annotation enrichment methods uncover functional modules of genes. This allows the modeller to concentrate on modelling the interactions between just those modules or the involved genes.

#### 5. Learning Algorithms for Network Inference

In general, network reconstruction is performed by applying a learning algorithm that fits the output of the mathematical model to the provided experimental data. The choice of an appropriate learning algorithm is mainly influenced by the selected model architecture (see Section 3) as well as by the quality and the quantity of the available data. Furthermore, if prior knowledge about gene regulatory interactions is available, the learning algorithm should be able to incorporate this knowledge into the final model (Section 6).

In network inference, two tasks can be distinguished: (1) the estimation of the model structure and (2) the estimation of the model parameters. Structure optimization corresponds to the problem of finding the network connectivity or topology that best explains the observed data and that simultaneously fulfils constraints representing the available knowledge, e.g. that takes the network-sparseness requirement into account (van Someren et al., 2001; Filkov, 2005). Parameter estimation concerns the problem of identifying the corresponding model parameters once a



**Fig. 5.** Overview of structure optimization methods that can be applied in particular for the learning of linear differential/difference equation models.

model structure is given (Section 5.1). To capture the network sparseness, most network inference approaches try to reduce the in-degree for each node. In many of these approaches the structure is optimized explicitly and parameter optimization becomes an embedded task of structure optimization (see Section 5.2.1). Alternatively, there exist several approaches where the structure is implicitly determined during parameter estimation (Section 5.2.2).

The estimation of other systems biology models, i.e. metabolic networks and signal transduction networks, is characterised by a mainly knowledge-driven determination of the model structure. Here, the focus lies on parameter estimation methods that surmount inherent ill-conditioning and multi-modality (Rodriguez-Fernandez et al., 2006; Moles et al., 2003). In contrast, in GRN inference a single node function has usually few parameters and nonlinearity is taken into account only in rare cases (e.g. in S-system models). Therefore, the decisive problem is the solution of the structure optimization problem. Note that the main focus of this section is on learning algorithms for differential and difference equation systems.

### 5.1. Parameter Optimization

The optimization of the parameters of a model is connected with the chosen model architecture and the scoring function that has to be optimized. The scoring function always contains a term quantifying the fit of the predicted model outputs to the gene expression data, also referred to as data-fit. Dependent on the assumed noise distribution, measures for this criterion are, e.g. the sum of squared errors (e.g. Mjolsness et al., 2000; Yeung et al., 2002; van Someren et al., 2006) or the maximum likelihood function. For each type of model architecture a large number of standard parameter optimization techniques are available, e.g. as presented Polisetti et al. (2006) for Generalized Mass Action models and by Vilela et al. (2008) for S-systems.

### 5.2. Structure Optimization

Deriving the model structure or the connectivity between the nodes is a challenging combinatorial optimization problem. For each node function the most likely combination of regulators has to be found. The total number of possible combinations for each node function is  $2^N - 1$ , where  $N$  is the number of nodes in the network. For a relatively small network with  $N = 20$  nodes the total number of possible regulatory combinations is about 1,000,000 for each node.

Consequently, even for small networks it is an impractical task to test all possible network structures. However, the number can be significantly decreased by the assumption of limited connectivity between the genes. If, e.g. the number of regulators is restricted to four and  $N = 20$ , then only 6195 regulatory combinations are possible. In this case, one might test all combinations by an exhaustive (brute-force) search.

A general rule in network reconstruction is that as the connectivity of the network increases, the model will better fit the data. However, several difficulties are concerned with higher-connected networks. First of all, genes are assumed to be regulated by a limited number of regulators (Arnone and Davidson, 1997; see Section 6.1). Secondly, the reliability of the parameter estimation deteriorates when the number of parameters increases. Due to the dimensionality problem (i.e. many parameters and few data), many network structures of sufficiently high connectivity can describe the same data equally well (Krishnan et al., 2007). Consequently, it is difficult to reliably determine which of these network structures is the best, making the inference results not robust. Therefore, a compromise between model quality and model complexity has to be found, which is known as the bias-variance trade-off. A general overview of structure optimization methods applied in experimental modelling is given in Nelles (2001).

As a remark, some network inference methods are characterised by the decomposition of the overall structure optimization problem into separate optimization steps. Then, in each step, the most likely regulators of a single gene have to be found.

#### 5.2.1. Explicit Structure Optimization

Explicit structure optimization methods examine GRN models with different topology and compare them by means of their scoring function (from Section 5.1). The scoring function is often augmented with a model complexity term aimed to prevent data over-fitting and ensure network sparseness. Such scoring functions were introduced for different network modelling formalisms, for instance BNs. Here, approximations such as the Bayesian Information Criterion (BIC) score are commonly used to assess the degree to which the resulting structure explains the data, while at the same time avoiding overtraining by penalizing the complexity (number of parameters) of the model.

Following a given strategy, interactions are added and removed trying to obtain a structure with a better score. Since testing all possible combinations of interactions can only be performed for very small networks, different structure optimization strategies exist that systematically search in the space of possible solutions (i.e. net-

work structures). As shown in Fig. 5, explicit structure optimization strategies comprise the simple testing of all possible combinations, heuristic search methods, evolutionary algorithms and simulated annealing (van Someren et al., 2001). Finally, given the modelling architecture and a network structure optimization strategy, a model can be estimated from the given data. The estimation might be supported by prior knowledge and additional types of data, e.g. by adapting the scoring function (see Section 6).

In case of very small networks or strong restrictions, e.g. limiting the number of regulators per gene, all possible combinations can be tested. For instance, Chen et al. (1999) suggested to simply test all combinatorial choices that have at most  $k$  regulators for a linear differential equation system. This brute-force strategy is often applied for inferring Boolean networks, too (e.g. Akutsu et al., 1999; Martin et al., 2007).

Heuristic search algorithms apply rules-of-thumb or guesses to guide the search in direction towards plausible solutions (most likely solutions first). Well known heuristic search techniques are, e.g. best first search, beam search and hill-climbing. In GRN inference, search algorithms might start from an initial topology and either add or remove interactions or reverse the direction of causality. Three types of main search strategies can be distinguished: forward selection (growing), backward elimination (pruning) and stepwise selection. Forward selection methods start with a simple model (e.g. without interactions) and add the most significant interactions until a stopping criterion is met. Alternatively, backward elimination methods start with a fully connected model and remove the least significant interactions until a stopping-criterion is met. Stepwise selection methods combine forward selection and backward elimination.

Note that similar heuristic structure learning algorithms are used for the inference of different model architectures, e.g. in BN inference to identify the most probable structure of a GRN learned from data (Heckerman, 1996; Needham et al., 2007). For the inference of BN models the REVEAL algorithm introduced by Liang et al. (1998) applies a forward selection algorithm, where subsequently all input pair combinations with  $k = 1, 2, 3, \dots$  regulators are examined. In Weaver et al. (1999) an advanced backward elimination strategy was suggested that removes recursively the interactions with the smallest parameter values, whereas the parameters are re-estimated in each iteration. Chen et al. (2001) proposed an inference method for an information theoretical model where first, putative pair-wise interactions are derived by correlating peaks in the time-series data and afterwards, less important interactions are eliminated. The NetGenerator algorithm (Toepfer et al., 2007; Guthke et al., 2005) combines forward selection and backward elimination to fit a system of linear or non-linear differential equations. Thereby, in order to avoid over-fitting (and to obtain a sparse network model), the inclusion or removal of interactions was tied to specific conditions, e.g. (i) an increase in model complexity must lead to a considerably improved model fit and (ii) the number of model interactions must not exceed a predefined limit.

The Inferelator inference method proposed by Bonneau et al. (2006) is based on a more complex differential equation system. In contrast to other approaches, the combination of regulators is not restricted to simple weighted sums. A special encoding of TF interactions allows to accommodate combinatorial logic (AND, OR, XOR) into the model. Here, this encoding was restricted to pair-wise interactions. Then, to fit a model for each gene, a combination of explicit and implicit structure optimization is performed: explicit structure optimization is utilised to get a selection of potential single TFs and pair-wise interactions. Implicit structure optimization selects the best combination for the final model using LASSO (see Section 3.3).

Different heuristic search strategies and genetic algorithms for the inference of GRNs using artificial data were compared by van Someren et al. (2001).

### 5.2.2. Implicit Structure Optimization

Implicit structure optimization is reached by optimising the model parameters using an extended scoring function. In addition to a model fitting term this extended scoring function includes a model complexity term which directly penalises the number of network interactions. This is also known as a form of regularisation. Regularisation reduces the effective number of parameters for each node function while the nominal number of optimized parameters corresponds to the total number of possible regulators. During parameter optimization the model is adopted to the measured data, while parameters not required for fitting are driven to zero. In consequence, a sparsely connected network results. Different implicit structure optimization methods can be distinguished with respect to the applied regularisation technique. Mjolsness et al. (2000) used a weight decay term that penalises the sum of the squared interactions weights within a non-linear differential equation system. A similar approach is proposed by van Someren et al. (2006). Their LARNA method (Least Absolute Regression Network Algorithm) minimises the sum of the absolute weights of a linear difference equation model. In Yeung et al. (2002) a two step procedure is applied to infer linear differential equations systems including SVD and subsequent robust regression.

## 6. Integration of Diverse Biological Information

As mentioned throughout this review, the inference of a large-scale GRN is complicated due to the combinatorial nature of the task and the limitations of the available data. Therefore, the use of prior knowledge and biologically plausible assumptions with respect to the model structure is essential to support the reverse engineering process. In addition, information from alternative experiments, various databases as well as from the scientific literature itself should be incorporated.

### 6.1. General Network Properties and Modelling Constraints

Several general properties of GRNs can be used for network reconstruction, including sparseness, scale-freeness, enriched network motifs and modularity. The most common and important design rule for modelling gene networks is that their topology should be sparse. Sparseness reflects the fact that genes are regulated only by a limited number of genes (Arnone and Davidson, 1997). Note that the term ‘sparse’ stands for limited regulatory inputs per gene, thus a low in-degree is desired. However, some so-called master genes may control a large part of the entire network, thus the out-degree is unrestricted. Enforcing the sparseness property during network identification has the benefit that it significantly reduces the number of model parameters to be estimated and consequently improves the quality of network inference. Techniques to constrain the number of regulators per gene are covered in Section 5.2. For instance, when scoring candidate models during structure optimization one might use scores that have a measure of how well the model fits the data, and a penalty term to penalise model complexity. Note that a drawback of limiting the number of edges in the network is that one may miss redundant paths in the network such as feed-forward loops.

Several studies have shown that the distribution of node degrees in biological networks often tends to have the form of a power law (Jeong et al., 2000; Bork et al., 2004), i.e. the fraction  $P(k)$  of nodes in the network having  $k$  connections goes as  $P(k) \sim k^{-\gamma}$ , where  $\gamma$  is a constant. In these, so-called scale-free networks, most of the genes are sparsely connected, while a few are very highly connected. Scale-freeness ensures the performance and robustness of networks with respect to random topological changes and is therefore an organising principle of biological structures (Jeong et al., 2000).

Not surprisingly, large-scale GRN models (information theory models) inferred from human gene expression data also demonstrate scale-free structures (Jordan et al., 2004; Basso et al., 2005). As scale-freeness is a stronger assumption than sparseness, it seems reasonable to utilize this property as a modelling constraint. Scale-freeness has been implemented by Chen et al. (2008) for a method that infers undirected edges based on a thresholded ranking of the most correlating genes by specifying whether a node is a core node or a periphery node. It came to our attention that at least one group considers scale-freeness during inference of dynamic GRNs (Westra, 2008). They first introduce a measure that compares how well the degree-distribution of a difference equation network model fits a perfect scale-free network. Then, they iteratively estimate the model parameters while maximizing this measure and optimizing  $\gamma$  until a convergence criterion is met. Alternatively, the concept of scale-freeness can be taken into account indirectly by limiting the number of candidate regulators in the network. Pre-defining known (and putative) TFs as regulators is a widely used approach to limit the model search space (e.g. Chen et al., 2001; Segal et al., 2003; Bonneau et al., 2006). However, one should be aware that the expression level of a TF does not necessarily reflect its activity.

Another property of natural regulatory networks is that they are highly structured. The low-dimensional connection structures in these networks follow regular hierarchies. This facilitates the decomposition of biological networks into basic recurring modular components that consist of only a few genes, so-called network motifs (Shen-Orr et al., 2002; Lee et al., 2002). Consequently, regulatory network motifs open the way to structured model identification. However, the use of such structural motifs is still under discussion. For instance, the handling of feedback loops is diverse and depends on the biological problem. Some authors reconstruct networks that are restricted to a hierarchical structure, e.g. Hartemink et al. (2002) using a BN formalism, whereas others forbid short loops. For example, ARACNE aims to remove indirect interactions from the inferred network. Thereby, if triplets of genes are fully connected, the edge with the weakest statistical relevance will be eliminated (Margolin et al., 2006). Apart from this, many reverse engineering algorithms are completely unrestricted to allow even short positive or negative feedback loops within the system (e.g. Liang et al., 1998; van Someren et al., 2002b).

Modularity is also an important property of GRNs. It is evident that genes share functionality and often act together, thus appearing to have a decentralised, redundant organisation. This property is well supported by the common occurrence of clusters of strongly co-expressed genes and correspondingly strong functional enrichment. The concept of modularity is important for the reconstruction of GRNs as it allows to tackle the data insufficiency problem. Therefore, a widely used approach is to group genes based on functional similarities or similar expression patterns (see Section 4.2) and then to model the regulatory interactions between those modules to get a higher-level view of gene regulatory mechanisms (e.g. Segal et al., 2003; Bar-Joseph et al., 2003; Guthke et al., 2005).

## 6.2. Integration of Heterogeneous Data

Many techniques have been proposed to identify GRNs from transcriptome data (e.g. obtained by DNA microarray experiments). Some authors derived dynamic network models from time-course gene expression data, e.g. D'haeseleer et al. (1999), van Someren et al. (2002b), Guthke et al. (2005). Others have utilized static expression data for network inference. For instance, in the study of Rung et al. (2002) an information theoretical model of the GRN of yeast was reconstructed from expression data of 274 different single gene deletion mutants. Further groups used both steady-

state and temporal measurements to compute hypothetical GRNs, e.g. of *Halobacterium* (Bonneau et al., 2006) and *E. coli* (Faith et al., 2007).

Although, DNA microarray data are widely used in the field of network inference, the reconstruction of GRNs using microarray data alone is inherently bounded as the information content of such data is limited by technical and biological factors. Therefore, more sophisticated methods have been developed to reconstruct the structure and dynamics of GRNs more reliably by incorporating other kinds of biological information. For instance, information on molecular interactions is accessible in many ways (see Section 2.1; Table 1) and thus can augment GRN modelling. Prior knowledge and additional large-scale experimental data also facilitate the reconstruction of more mechanistic models. Note that the prior knowledge utilized must suit the given data and the scientific question of the study.

An integrative learning strategy often consists of two steps. First, a template of the network is built using various levels of additional information, e.g. from databases and the literature. This template represents a supposition of the real underlying network topology. Second, an inference strategy is applied that fits the model to the data while taking the template into account. The template information can be incorporated into the network inference process, e.g. in Bayesian frameworks by appropriately setting prior probabilities of the network structure. A more general approach is to let the template adapt the cost function or to simply use the template to constrain explicit search methods.

A BN is a good representation of the combination of prior knowledge and data because it reflects both causal and probabilistic semantics. More exactly, the integration of biological knowledge can be realised by inferring the model in a maximum a posteriori sense. Formally, the probability distribution for a model  $\theta$  given data  $D$  and background knowledge  $\xi$  is according to the Bayes' theorem:

$$p(\theta|D, \xi) = p(\theta|\xi)p(D|\theta, \xi)/p(D|\xi).$$

The probability distributions  $p(\theta|\xi)$  and  $p(\theta|D, \xi)$  are commonly referred to as the prior and posterior for  $\theta$ , respectively.  $p(D|\theta, \xi)$  is the likelihood of the "data given model", i.e. describes the fitness of a model to the data, and we assume here that  $D$  and  $\xi$  are independent. If prior knowledge is available, the prior defines a function that measures the agreement between a given network and the biological prior knowledge (template) that we have at our disposal. There are many types of priors that may be used, and there is much debate about the respective choice. Heckerman (1996), Needham et al. (2007) as well as Werhli and Husmeier (2007) are excellent tutorials on learning with BNs using prior knowledge. Commonly used heuristics to learn BNs (i.e. to identify the most probable GRN structure) are covered in Section 5.

As shown for BNs inferred from synthetic data, the integration of prior knowledge about the network topology increases the network reconstruction accuracy (Le et al., 2004; Geier et al., 2007). As a concrete example, Imoto et al. (2003) derive GRNs from microarray gene expression data, and use biological knowledge (regulatory interactions from the Yeast Proteome Database) to effectively favour biologically relevant network structures. Thereby, according to the BN framework explained before, the fitness of each model to the data was first measured and subsequently biological knowledge was input in the form of a prior probability for structures (in this case expressed in terms of an energy function). Then, the posterior probability for the proposed GRN was the product of the fitness and the prior probability of the structure. With this in mind, TF-DNA binding data was applied complementary to DNA microarray data. In the work of Hartemink et al. (2002), TF-DNA interactions found by ChIP analysis were incorporated into the modelling of a network of 32 selected yeast genes. Thereby, BN models that failed to include an edge where the location data suggested one were eliminated from consideration *a priori* (by setting  $p(\theta|\xi)=0$ ). In a later work

by Bernard and Hartemink (2005), these constraints were relaxed. Here, edges for which location data indicates TF–DNA interactions were more likely though not forcibly included in the model, considering that the prior knowledge is not infallible. Similarly, TF–DNA interactions predicted by analysing promoter DNA sequences for TFBS were used in combination with gene expression data (Tamada et al., 2003; Jensen et al., 2007). Information on protein–protein interactions have also been used to refine GRNs estimated from expression data (Nariai et al., 2004). Here, the biological implications of protein–protein interactions were incorporated in the learning scheme by adding nodes representing protein complexes when the resulting BN structure is better suited to reflect the data.

The application of BNs for knowledge supported network inference is an active field of research. However, analogously, the incorporation of prior knowledge can be realised within different inference architectures by appropriately setting a model fitness scoring function (e.g. as a weighted sum of data-fit and template-fit). Variants of this approach have been proposed for Boolean networks (Birkmeier, 2006), linear difference equation models (Yong-A-Poi, 2008; Koczan et al., 2008) and non-linear differential equation models (Spieth et al., 2005). For instance, a linear difference equation model can be inferred using prior knowledge by an adaptation of the LASSO method. The LASSO fits the model to the data in a least-squares sense subject to  $\sum_j |\beta_j| \leq s$ ,  $s > 0$ . Because of the nature of this constraint it tends to produce some coefficients  $\beta_j$  (model parameters) that are exactly zero and hence gives sparse, interpretable models (the lower  $s$ , the sparser the resulting model). Now, a template (i.e. prior knowledge) can be used by assigning different weights to the coefficients:  $\sum_j \tilde{w}_j |\beta_j| \leq s$ ,  $s > 0$ . Thereby, a relatively low weight  $\tilde{w}_j$  provokes that the edge corresponding to  $\beta_j$  is preferred to be in the final model. This concept was applied to integrate human microarray data with gene regulatory interactions obtained by text mining by Yong-A-Poi (2008) and Koczan et al. (2008), in which  $\tilde{w}_j$  was defined as a constant and as function of  $\beta_j$ , respectively.

However, whenever heterogeneous data and additional information from the literature are incorporated into the inference process, one has to keep in mind that the quality of the inferred models always depends on the quality and completeness of this additional/prior knowledge. Today, *S. cerevisiae* is one of the best-studied model organisms. It is hence not surprising that a lot of GRN modelling studies focused on this organism (Hartemink et al., 2002; Rung et al., 2002; Bar-Joseph et al., 2003; Segal et al., 2003; Imoto et al., 2003; Tamada et al., 2003; Nariai et al., 2004; Bernard and Hartemink, 2005; Jensen et al., 2007; Larsen et al., 2007). As more and more specific information becomes available, the inference of (dynamic) network models supported by diverse sources of biological knowledge will be more frequently carried out for other organisms as well. A so far underexplored topic is the trade-off between data-fit and confidence in the prior knowledge, i.e. the difficulty to conveniently set the confidence associated with the prior knowledge relative to the expected noise in the data.

## 7. Network Validation and Assessment of the Network Inference Methods

Network validation consists of assessing the quality of an inferred model with available knowledge. For quantitative validation of an inferred GRN, it is necessary to employ a scoring methodology that evaluates the model with respect to (a) information already used to generate the model (internal validation) and (b) information independent from the information used to reconstruct the network (external validation).

### 7.1. Scoring Methodology

In general, the quality of a GRN model can be evaluated by the answer to one or both of the two questions:

- (a) Does the model correctly predict the behaviours of the GRN?
- (b) Does the model represent the true structure of the system?

Answering the first question, one compares the simulated behaviour of the model system with the measured or observed behaviour of the real system. This can be quantified by cost functions that are also used for model optimization as discussed in Section 5. Answering the second question, one needs at least partial knowledge about the true interactions, which is generally incomplete, uncertain or difficult to obtain (especially when modelling a network of gene modules) in practice. For the assessment of network inference methods one might overcome this problem by employing synthetic data generated from artificial networks (see Section 7.4). Supposing that a representation of the true structure of the network is known or can be obtained (e.g. by direct experimental verification or database search), the predicted network structure can be compared to this ‘true network’ based on a variety of performance measures. To this end, the number of truly (T) and falsely (F) predicted regulatory edges is counted, and the presence or absence of interactions between nodes is referred to as positive (P) or negative (N) respectively. Now, the following numbers can be defined:

- TP = the number of true positives, i.e. the number of correctly inferred edges;
- FP = the number of false positives, i.e. the number of inferred edges that are incorrect;
- TN = the number of true negatives, i.e. the number of missing edges in the inferred network that are also missing in the true network;
- FN = the number of false negatives, i.e. the number of missing edges in the inferred network that are an edge in the true network.

Note that this nomenclature is based on a binary classification of edges, i.e. does an edge occur in the network or not. This approach is sufficient in most cases as it can be applied on both directed and undirected networks. However, to distinguish between inhibiting and activating effects, similar counts could be defined for the three classes of ‘activation’, ‘inhibition’ or ‘no effect’. For instance, the situation of having “inferred an activation” while an inhibition was expected might be counted as a false positive prediction. In rare cases, one might even want to assess how close the strength of interactions was inferred (e.g. using the Euclidean metric on the expected and inferred continuous model parameters).

Based on the previously defined binary counts, performance scores can be computed. The ‘recall’ or ‘sensitivity’ is defined by  $TP/(TP + FN)$  and denotes the fraction of correctly identified interactions in relation to the number of expected interactions. ‘Precision’ is determined by  $TP/(TP + FP)$  and denotes the fraction of correctly identified interactions out of all predicted interactions. ‘Specificity’ computed by  $TN/(TN + FP)$  measures the proportion of non-existing edges (number of potential edges – number of inferred edges) which are correctly identified. Further commonly used scores are the false positive rate ( $=FPR = 1 - specificity$ ) and the false discovery rate ( $=FDR = 1 - precision$ ). Note that each of these scores is calculated only from two numbers out of {FN, FP, TP, TN}, i.e. each score is hardly informative when used alone. For instance, an inferred fully-connected network will result in a recall equal to 1, but is obviously not biologically meaningful.

Typically, when inferring a GRN one (a) has a ranking on the edges reflecting the reliability of the predictions (e.g. an ordering on pair-wise computed correlation coefficients of an information

theory model) or (b) can adjust the parameters of the inference learning scheme to obtain networks of low, moderate and high connectivity. Then, the performance of the network inference algorithm can be visualised as a precision-versus-recall curve (PRC-curve). The curve results from increasing the number of edges predicted following (a) or (b). Alternatively, a similar curve results when visualising recall versus FPR (receiver operating characteristic or simply ROC-curve). Both, PRC-curve and ROC-curve have advantages and disadvantages, thus they are usually used together to evaluate the performance of different inference techniques (Soranzo et al., 2007; Stolovitzky et al., 2007). In general, the ROC analysis is only valid for the binary classification problem indicated above, but allows to directly compare the inference quality against a random prediction by calculating the area under the curve (AUC), which is often used as a single metric in benchmark tests. An AUC(ROC) close to 0.5 corresponds to a random forecast, AUC(ROC)<0.7 is considered poor, AUC(ROC)>0.8 good (Soranzo et al., 2007). However, since GRNs are sparse, FP might far exceed TP. Thus, specificity ( $1 - \text{FPR}$ ) which is used in ROC analysis, is inappropriate as even small deviations from a value of 1 will result in large FP numbers. For this reason, the PRC-curve can be a more useful component for GRN performance evaluation.

## 7.2. Internal Validation

In statistics, there are different resampling techniques to evaluate the generalization performance or robustness of a model, e.g. subsampling, bootstrapping and perturbation. Subsampling, e.g. cross-validation, and bootstrapping are based on splitting the available data into training and test data sets. In  $k$ -fold cross-validation, the data set is partitioned into  $k$  subsamples. A single subsample is retained as the test data set, and the remaining  $k - 1$  subsamples are used for training. Subsampling and bootstrapping are not well suited for time series data (since splitting such data makes little sense). Instead, the effect of measurement noise on the inferred model might be assessed by repeated network inference on randomly perturbed data (D'haeseleer et al., 2000; Guthke et al., 2005). Thereby, the noise added to the measured data should be of the order of magnitude of the measurement noise or biological variability (Moeller and Radke, 2006).

## 7.3. External Validation: Knowledge- and Experiment-based Validation

The internal model validation may be insufficient because the presumptions that underlie the chosen modelling architecture (Section 3) and modelled components (Section 4) may oversimplify the true complexity in GRNs. In addition, the available data is mostly inadequate with respect to the data requirements for large-scale models (Section 2.3). Often, the inference result is not unique, i.e. some model elements cannot be identified. Therefore, model predictions should be checked by data, information and observations that were not used for modelling. Subjects for external validation are knowledge available from literature or databases, and data from experiments possibly initiated in response to the modelling. By using such additional information, an assessment of the network reconstruction is possible by the scores explained in Section 7.1. Exemplarily, in the work of van Someren et al. (2006) knowledge-based validation employing text mining information was used to assess and compare diverse network inference methods. Recently, the elegant concept to integrate half of available prior knowledge into the network inference and subsequently validate the model on the remaining knowledge was addressed by Yong-A-Poi (2008). However, knowledge-based model validation is unsuited to validate novel insights of the GRN model. To set an example, Perkins et al. (2006) compared the behaviour of five

models inferred from data and two models found in the literature describing early *Drosophila melanogaster* development. Interestingly, some inferred relationships were found to be inconsistent with standard textbook models, thus experimental validation is inevitable.

## 7.4. Assessment of The Network Inference Methods

The assessment of GRN inference algorithms requires benchmark data sets for which the underlying network is known. However, experimental (gold standard) data sets with the corresponding 'complete' knowledge of the network structure are hardly available, even if there is ongoing work. Hence, there is a need to generate synthetic data that allow for thorough testing of learning algorithms in a reproducible manner. The inherent weakness of such approach is that the performance of an inference strategy would strongly rely on the model used to construct the artificial data. Zak et al. (2001), Mendes et al. (2003) and others proposed models and tools for generation of synthetic data that include rates of transcription and mRNA degradation. Using synthetic data from models introduced by Zak et al. (2001) in a slightly modified form and by analyzing ROC-curves, Husmeier (2003) demonstrated how the performance of network inference by employing DBNs depends on the reliability of prior assumptions, the size of the training set and the number of sampling points. Synthetic gene expression data from *in silico* 'experiments' simulated by models similar to the model from Mendes et al. (2003) were used by Yeung et al. (2002) to introduce a novel algorithm that combines SVD with robust regression. They concluded from their analyses that the number of sample points needed to recover a sparsely connected network scales logarithmically with the size of the network. Synthetic data were also applied by Geier et al. (2007) to compare the performance of DBNs and linear regression with variable selection based on  $F$ -statistics. They used synthetic data simulated by a non-linear model according to Mendes et al. (2003) representing 10 TFs and 20 other genes to study specific perturbations of the GRN in the form of TF knock-outs and the use of prior knowledge.

Faith et al. (2007) applied the CLR algorithm (see Section 3.1) and compared its performance with other popular inference strategies (ARACNE, RELNET, linear regression networks) on a compendium of 445 DNA microarray experiments for *E. coli*. When evaluated against known regulatory interactions from RegulonDB, both CLR and RElevance NETworks reach high precisions, but CLR attains almost twice the sensitivity of RELNET at some levels of precision. The algorithms NIR, MNI and TSNI (see Section 3.3.1) were benchmarked by Bansal et al. (2006) on a synthetic data set. They showed that the reverse engineering tools MNI and TSNI are not well suited for inferring large-scale networks, but rather for identification of the targets of a perturbation. In a later work Bansal et al. (2007) evaluated public software tools (ARACNE, BANJO and NIR—see Section 3) using both synthetic and experimental microarray data with the following conclusions: ARACNE performed well for steady-state data, but was not suited for the analysis of short time-series data. NIR worked very well for steady-state data, but required knowledge on the genes that have been perturbed directly. BANJO required a large number of data points, but when this condition was met, it performed comparably to the other methods.

Noteworthy, the Dialogue on Reverse Engineering Assessment Methods (DREAM) is fostering a concerted effort by computational and experimental biologists to understand the limitations and strengths of techniques for inferring networks from high-throughput data through network inference challenges. Thereby, they aim to create what seems to be a suitable set of gold standards for network inference assessment by providing curated data sets to the community and defining common evaluation metrics (Stolovitzky et al., 2007). A recent example of the DREAM initiative

**Table 2** Overview of selected GRN inference approaches found in the literature. Shown are the type and amount of the gene expression data used in each work as well as the methods used to extract the relevant features from this data. The column '#Nodes' lists the number of nodes (genes or clusters) actually considered in the respective network model. Details on the applied inference techniques are given in the next column. The column 'Data integration' indicates whether or not additional data was used to support the inference process. The column 'Constraints' shows which of the general modelling constraints were used: SP—sparseness (i.e. the network structure is constrained to be sparse); RL—indicates whether or not the number of regulators is limited, e.g. to previously known TFs (which implies sparseness); and CS—indicates whether or not expression is thought to change smoothly over time (i.e. additional time points were estimated by interpolation to 'increase' the amount of data). The column furthest right shows methods that were applied to validate the inference results. Further abbreviations used in this table:  
#Genes—number of genes measured; #Observ.—number of observations; Oligon.—oligonucleotide; ma.—microarray; tp.—time points; exp.—experiments; expr.—expression; stat.—statistical.

Reference	Gene expression data			Feature selection			Inference technique			Validation method					
	Author (year)	Type	Organism	#Genes	#Observ.	Filtering	Clustering	#Nodes	Model scheme	Learning algorithm	Data integration	Constr.	SP	RL	CS
Dhaeseleer et al. (1999)	RTQ-PCR	Rat		65	Time-series (28 tp.)	—	—	65	Linear difference	Least squares	—	—	✓	—	—
Chen et al. (2001)	Oligon. ma.	Yeast		6,601	Time-series (17 tp.)	Excluding low expr.	Hierarchical clustering	308	Information-theoretical	Stepwise (simulated annealing)	—	✓	✓	—	
Hartemink et al. (2002)	Oligon. ma.	Yeast		6,135	Static (320 exp.)	Knowledge-driven	—	32	Bayesian network	Stepwise (simulated annealing)	TF-DNA binding data	✓	—	—	
Inmoto et al. (2003)	cDNA ma.	Yeast		6,000	Static (100 exp.)	Knowledge-driven	—	36	Bayesian network	Stepwise (hill climbing)	Databases, literature	✓	—	—	
Tamada et al. (2003)	cDNA ma.	Yeast		5,871	Static (100 exp.)	Knowledge-driven	—	124	Bayesian network	Maximum likelihood re-estimations	DNA sequence (motif search)	✓	—	—	
Narai et al. (2004)	cDNA ma.	Yeast		6,178	Time-series (69 tp.)	Knowledge-driven	—	99	Bayesian network	Stepwise (hill climbing)	Protein-protein interaction data	✓	—	External (KEGG database)	
Basso et al. (2005)	Oligon. ma.	Human		~10,000	Static (336 exp.)	—	—	~10,000	Information-theoretical	Brute force	—	✓	—	Experimental	
Bernard and Hartemink (2005)	cDNA ma.	Yeast		6,178	Time-series (69 tp.)	Knowledge-driven	—	25	Dynamic Bayesian network	Stepwise (simulated annealing)	TF-DNA binding data	✓	—	—	
Guthrie et al. (2005)	cDNA ma.	Human		7,619	Time-series (5 tp.)	Fold-criterion	Fuzzy c-means clustering	6	Linear differential equations	Stepwise	—	✓	—	Data-based (repeated perturbation)	
Kimura et al. (2005)	cDNA ma.	<i>T. thermophilus</i>		612	Time-series (14 tp.)	—	Hierarchical clustering	25	S-system model	Evolutionary algorithm	—	✓	—	—	
Bonneau et al. (2006)	Oligon. ma.	<i>Halobacterium</i>		~2,400	Mixed (268 exp.)	Excluding low expr.	Biclustering	531	Generalized linear difference equations	Bivariate selection prior LASSO	—	✓	✓	Data-based (cross-validation); Experimental	
van Someren et al. (2006)	cDNA ma.	Mouse		9,596	Time-series (5 tp.)	—	—	101	Linear difference equations	LASSO	—	✓	✓	External (literature)	
Faith et al. (2007)	Oligon. ma.	<i>E. coli</i>		4,345	Mixed (445 exp.)	—	—	4,345	Information-theoretical	Brute force	—	✓	✓	Experimental	
Martin et al. (2007)	Oligon. ma.	Mouse		~34,000	Time-series (12 tp.)	Excluding low expr.; stat. significance	k-Means clustering	12	Boolean network	Brute force	—	✓	—	—	
Koczan et al. (2008)	RTQ-PCR	Human		20	Time-series (19 × 3 tp.)	—	—	20	Linear difference equations	LASSO	Databases, literature	✓	✓	—	

is the five-gene network challenge. In this challenge, they provide expression data obtained from a synthetic 5-gene network in yeast, i.e. a network by human design that was transfected into an *in vivo* model organism. This allows the inference of a GRN for which the true network structure is known.

## 8. Conclusions

Discovering structures and dynamics of GRNs based on large-scale data represents a major challenge in systems biology. There is a vast variety of data and network types, inference methods as well as evaluation metrics for network inference. Even if the different model architectures rely on completely different mathematical formalisms, all models can be interpreted as networks of interacting nodes. Nodes represent molecular entities such as genes and proteins, or functional modules, whereas edges correspond to regulatory interactions and other relations between those nodes. Due to limitations in the amount and quality of available data and the corresponding computational efforts, network inference methods require simplifications such as linearization, discretization or aggregation of compounds to modules. The usefulness of a GRN inference method mainly depends on both the intended application of identified networks and the data at hand.

**Table 2** provides an overview of the characteristics of different reverse engineering studies, covering the used data, feature selection methods, inference techniques, constraints and validation methods.

### 8.1. The Purpose

Mathematical models can be used in two different ways (see also Gardner and Faith, 2005): first, the use of ‘mechanistic’ network models aims to identify true molecular interactions. These include protein–DNA interactions, in particular the interactions of TFs with binding sites of their target genes, as well as protein–protein and protein–ligand interactions forming signalling pathways. Due to the vast amount of molecules in cells, it is necessary to mention that such reverse engineering approaches do not claim to recover the totality of connections in a biological network but rather reveal interactions that are highly significant under defined (experimental) conditions.

Second, so-called ‘influence’ network models generally reflect global properties of a system’s behaviour. Influence networks relate the expression of one gene or a group of genes (module) to the expression of another gene or module. Using the influence approach, true molecular interactions are described rather implicitly. Therefore, influence models can be difficult to interpret and also difficult to integrate or extend using further information. Solely analysing gene expression data allows to infer influence networks of gene-to-gene interactions. Though, the integration of prior knowledge as well as the use of additional experimental data can lead to network models whose edges might be interpreted more mechanistically in terms of molecular interactions.

### 8.2. The Data

Data obtained from DNA microarray monitoring of gene expression are the most common type of data used to reverse engineer GRNs. Other less mature high-throughput techniques are emerging and improving at a rapid pace. However, with respect to data quality and quantity, no single measurement technique is capable of providing all necessary data for an error-free network inference. Deeper biological insight will be gained combining different types of information including measurements of transcript levels, proteins and small molecules, as well as interactome measurements. Considering network edges that are supported by more than one of

these data sets will further increase the chance to actually identify biologically relevant interrelations.

The identifiability of model structure and parameters depends on the chosen model architecture and the modelled features (see Section 4) as well as on the experimental design (e.g. the kind of intervention; see Sections 2.2 and 2.3). Perturbation experiments by environmental changes such as heat shock or starvation alter the behaviour of the system in a non-specific way, often initiating extensive changes in the cellular behaviour. Experiments that apply specific techniques of intervention, such as gene knock-out or RNA interference, are able to generate highly informative data for network inference. This has been impressively demonstrated for simple microorganisms, e.g. *Halobacterium* (Bonneau et al., 2006), *E. coli* (Faith et al., 2007) and *S. cerevisiae* (Lee et al., 2002).

The quantity and quality of data available today is in general insufficient to infer mechanistic networks on a genome-wide scale. Only a small portion of the actually existing interactions can be identified by current approaches. The higher the number of interacting compounds (genes, proteins, etc.) the higher the complexity of a corresponding network model and thus, a larger number of both state variables and model parameters is required.

### 8.3. The Integration of Diverse Biological Information

The dimensionality (data insufficiency) problem strongly impedes the modelling of GRNs. Hence, in order to obtain reliable inference results, it is important to carry out feature selection, to incorporate biologically motivated constraints (such as sparseness) and to combine diverse types of data (e.g. gene expression data and sequence information). While network sparseness is commonly postulated during inference and implemented by limiting the number of regulators per gene or in general penalising model complexity, the properties of scale-freeness and modular design of regulatory networks have just been recognised as additional modelling constraints. As shown, various data and information from scientific literature and biological databases can be used in combination with gene expression levels, e.g. genome sequence data (TF binding motifs), gene functional annotations, text-mining information, ChIP-on-chip data and protein–protein interaction data. We reviewed promising studies that integrate such diverse types of data during the reconstruction of (dynamic) GRN models. The incorporation of heterogeneous data and prior biological knowledge has been presented in particular for Bayesian networks and linear difference equation models. Facing limited amounts of experimental data, such a combined analyses of different types of biological information supports the inference process and thus allows to infer more exact and more interpretable models. The integration of multiple sources of heterogeneous data and prior biological knowledge will be one of the major focuses in future GRN research.

### 8.4. The Assessment of Network Inference Methods

Current efforts aim to understand individual strengths and weaknesses of various GRN inference methods by applying them to equal data sets. Such comparisons require an appropriate evaluation scheme to assess the success and correctness of network reconstruction. Generally, researchers apply so-called ‘synthetic networks’. Here, designed networks are thought to produce artificial data approximating real gene expression values. Data produced by synthetic networks may be used to address questions like: Which experiments and data types are best suited for a specific network inference method? For individual methods, which algorithm configuration works best? Obviously, models used to generate synthetic data cannot reflect the complexity of a real biological system. However, standards are still missing to evaluate different inference methods using real biological data.

Systems biological models are intended to assist biologists in generating assumptions for further research activities. Hypotheses generated by modelling can and should be experimentally tested. **Faith et al. (2007)**, for instance, tested and confirmed predicted interactions using ChIP. The predictive power of a GRN model inferred by **Bonneau et al. (2006)** was successfully verified using DNA microarray data which were not included in the data set used for network inference. The validation and interpretation of GRN models ideally goes in line with new knowledge and experimental data available for modelling, and thus a reiterative cycle between model construction and experimental validation can be formed. It is exciting to see, how the modelling of GRNs can be improved by advances in biotechnology and bioinformatics in the future.

### Acknowledgements

We thank the reviewers for helpful comments and we would like to thank Dr. Michael Pfaff, BioControl Jena GmbH, for his work and advice on the manuscript. This work has been supported by the German Federal Ministry of Education and Research (BMBF, grants no. 0313078D and 0313692D).

### References

- Akutsu, T., Miyano, S., Kuhara, S., 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 17–28.
- Arnone, M.I., Davidson, E.H., 1997. The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124, 1851–1864.
- Bansal, M., Gatta, G.D., di Bernardo, D., 2006. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* 22 (7), 815–822.
- Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D., 2007. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 122 (3), 78 (corrigendum 3).
- Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A., Gifford, D.K., 2003. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* 21, 1337–1342.
- Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A., 2005. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382–390.
- Bernard, A., Hartemink, A.J., 2005. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 459–470.
- Birkmeier, B., 2006. Integrating Prior Knowledge into the Fitness Function of an Evolutionary Algorithm for Deriving Gene Regulatory Networks (Master Thesis). University of Skövde, Sweden.
- Bonneau, R., Reiss, D.J., Shannon, P., Facciotti, M., Hood, L., Baliga, N.S., Thorsson, V., 2006. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* 7 (5), R36.
- Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I., Marcotte, E.M., 2004. Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* 14 (3), 292–299.
- Bornholdt, S., 2008. Boolean network models of cellular regulation: prospects and limitations. *J. R. Soc. Interf.* 5, S85–S94.
- Butte, A., Kohane, I., 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 418–429.
- Chen, T., He, H.L., Church, G.M., 1999. Modeling gene expression with differential equations. In: Proceeding of the Pacific Symposium on Biocomputing, vol. 4, pp. 29–40.
- Chen, T., Filkov, V., Skiena, S., 2001. Identifying gene regulatory networks from experimental data. *Parallel Comput.* 27 (1–2), 141–162.
- Chen, G., Larsen, P., Almasri, E., Dai, Y., 2008. Rank-based edge reconstruction for scale-free genetic regulatory networks. *BMC Bioinform.* 9, 75.
- Cho, K.-H., Choo, S.-M., Jung, S.H., Kim, J.-R., Choi, H.-S., Kim, J., 2007. Reverse engineering of gene regulatory networks. *IET Syst. Biol.* 1 (3), 149–163.
- Climescu-Haulica, A., Quirk, M.D., 2007. A stochastic differential equation model for transcriptional regulatory networks. *BMC Bioinform.* 8 (Suppl. 5), S4.
- De Jong, H., 2002. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67–103.
- D'haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R., 1999. Linear modeling of mRNA expression levels during CNS development and injury. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 41–52.
- D'haeseleer, P., Liang, S., Somogyi, R., 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16 (8), 707–726.
- di Bernardo, D., Thompson, M., Gardner, T., Chobot, S., Eastwood, E., Wojtovich, A., Elliott, S., Schaus, S., Collins, J., 2005. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 23 (3), 377–383.
- Dougherty, E.R., Barrera, J., Brun, M., Kim, S., Cesar, R.M., Chen, Y., Bittner, M., Trent, J.M., 2002. Inference from clustering with application to gene-expression microarrays. *J. Comput. Biol.* 9 (1), 105–126.
- Ernst, J., Vainas, O., Harbison, C.T., Simon, I., Bar-Joseph, Z., 2007. Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.* 3, 74.
- Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., Gardner, T.S., 2007. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5 (1), e8.
- Filkov, V., 2005. Identifying gene regulatory networks from gene expression data. In: Aluru (Ed.), *Handbook of Computational Molecular Biology*. CRC Press, Chapman & Hall, pp. 27.1–27.29.
- Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., Mello, C.C., 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391, 806–811.
- Friedman, N., Linial, M., Nachman, I., Peer, D., 2000. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7 (6), 601–620.
- Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J., 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301, 102–105.
- Gardner, T.S., Faith, J.J., 2005. Reverse-engineering transcription control networks. *Phys. Life Rev.* 2, 65–88.
- Geier, F., Timmer, J., Fleck, C., 2007. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Syst. Biol.* 1, 11.
- Gibbons, F.D., Roth, F.P., 2002. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* 12 (10), 574–581.
- Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G., Kell, D.B., 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* 22, 245–252.
- Goutsias, J., Lee, N.H., 2007. Computational and experimental approaches for modeling gene regulatory networks. *Curr. Pharm. Des.* 13 (14), 1415–1436.
- Granzow, M., Berrar, D., Dubitzky, W., Schuster, A., Azuaje, F.J., Eils, R., 2001. Tumor classification by gene expression profiling: comparison and validation of five clustering methods. *SIGBio Newsletter Special Interest Group on Biomedical Computing of the ACM* 21, 16–22.
- 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 (8), 1626–1634.
- Guthke, R., Kniemeyer, O., Albrecht, D., Brakhage, A.A., Möller, U., 2007. Discovery of gene regulatory networks in *Aspergillus fumigatus*. *Lect. Notes Bioinform.* 4366, 22–41.
- Hartemink, A., Gifford, D., Jaakkola, T., Young, R., 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In: Proceeding of the Pacific Symposium on Biocomputing, vol. 6, pp. 422–433.
- Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A., 2002. Combining location and expression data for principled discovery of genetic regulatory network models. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 437–449.
- Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning*. Springer, New York.
- Heckerman, D., 1996. A Tutorial on Learning with Bayesian Networks. Microsoft Research Tech. Report, MSR-TR-95-06.
- Heinrich, R., Schuster, S., 1996. *The Regulation of Cellular Systems*. Chapman and Hall, 115 Fifth Avenue New York, NY 10003.
- Holter, N.S., Maritan, A., Cieplak, M., Fedoroff, N.V., Banavar, J.R., 2001. Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. U.S.A.* 98 (4), 1693–1698.
- Husmeier, D., 2003. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* 19 (17), 2271–2282.
- Ideker, T.E., Thorsson, V., Karp, R.M., 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 305–316.
- Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., Miyano, S., 2003. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. In: Proceeding of the 2nd IEEE Computer Society Bioinformatics Conference, pp. 104–113.
- Jensen, S.T., Chen, G., Stoeckert, C., 2007. Bayesian variable selection and data integration for biological regulatory networks. *Ann. Appl. Stat.* 1, 612–633.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.L., 2000. The large-scale organization of metabolic networks. *Nature* 407 (6804), 651–654.
- Jordan, I.K., Mariño-Ramírez, L., Wolf, Y.I., Koonin, E.V., 2004. Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.* 21 (11), 2058–2070.
- Kaern, M., Elston, T.C., Blake, W.J., Collins, J.J., 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6 (6), 451–464.
- Kauffman, S.A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* 22, 437–467.
- Kawasaki, E.S., 2006. The end of the microarray Tower of Babel: will universal standards lead the way? *J. Biomed. Tech.* 17 (3), 200–206.
- Khatari, P., Draghici, S., 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21 (18), 3587–3595.
- Kimura, S., Ide, K., Kashihara, A., Kano, M., Hatakeyama, M., Masui, R., Nakagawa, N., Yokoyama, S., Kuramitsu, S., Konagaya, A., 2005. Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics* 21 (7), 1154–1163.

- Koczan, D., Drynda, S., Hecker, M., Drynda, A., Guthke, R., Kekow, J., Thiesen, H.J., 2008. Molecular discrimination of responders and nonresponders to anti-TNF $\alpha$  therapy in rheumatoid arthritis by etanercept. *Arthritis Res. Ther.* 10 (3), R50.
- Krishnan, A., Giuliani, A., Tomita, M., 2007. Indeterminacy of reverse engineering of Gene Regulatory Networks: the curse of gene elasticity. *PLoS ONE* 2 (6), e562.
- Larsen, P., Almasri, E., Chen, G., Dai, Y., 2007. A statistical method to incorporate biological knowledge for generating testable novel gene regulatory interactions from microarray experiments. *BMC Bioinform.* 8, 317.
- Le, P.P., Bahl, A., Ungar, L.H., 2004. Using prior knowledge to improve genetic network reconstruction from microarray data. *Silico Biol.* 4 (3), 335–353.
- Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A., 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298 (5594), 799–804.
- Liang, S., Fuhrman, S., Somogyi, R., 1998. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 18–29.
- Ljung, L., 1999. System Identification—Theory for the User. Prentice Hall, Upper Saddle River, NJ.
- Margolin, A., Nemenman, I., Bassi, K., Wiggins, C., Stolovitzky, G., Favera, R., Califano, A., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform.* 7 (Suppl. 1), S7.
- Markowetz, F., Bloch, J., Spang, R., 2005. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics* 21 (21), 4026–4032.
- Markowetz, F., Spang, R., 2007. Inferring cellular networks—a review. *BMC Bioinform.* 8 (Suppl. 6), S5.
- Martin, S., Zhang, Z., Martino, A., Faulon, J.L., 2007. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics* 23 (7), 866–874.
- Mello, C.C., Conte Jr., D., 2004. Revealing the world of RNA interference. *Nature* 431, 338–342.
- Mendes, P., Sha, W., Ye, K., 2003. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* 19 (Suppl. 2), ii122–ii129.
- Mjolsness, E., Mann, T., Castano, R., Wold, B., 2000. From coexpression to coregulation: an approach to inferring transcriptional regulation among gene classes from large-scale expression data. In: Stoll, S.A., Leen, T.K., Müller, K.R. (Eds.), Advances in Neural Information Processing Systems, vol. 12. MIT Press, Cambridge, MA, pp. 928–934.
- Moeller, U., Radke, D., 2006. Performance of data resampling methods for robust class discovery based on clustering. *Intell. Data Anal.* 10 (2), 139–162.
- Moles, C.G., Mendes, P., Banga, J.R., 2003. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* 13 (11), 2467–2474.
- Mordelet, F., Vert, J.P., 2008. SIRENE: supervised inference of regulatory networks. *Bioinformatics* 24 (16), i76–82.
- Morgan, B.J.T., Roy, A.P.G., 1995. Non-uniqueness and inversions in cluster analysis. *Appl. Stat.* 44, 117–134.
- Nariai, N., Kim, S., Imoto, S., Miyano, S., 2004. Using protein–protein interactions for refining gene networks estimated from microarray data by Bayesian networks. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 336–347.
- Needham, C.J., Bradford, J.R., Bulpitt, A.J., Westhead, D.R., 2007. A primer on learning in Bayesian networks for computational biology. *PLoS Comput. Biol.* 3 (8), e129.
- Nelles, O., 2001. Nonlinear System Identification. Springer-Verlag, Berlin Heidelberg.
- Opgen-Rhein, R., Strimmer, K., 2007. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.* 1, 37.
- Pan, W., 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18 (4), 546–554.
- Pandey, A., Mann, M., 2000. Proteomics to study genes and genomes. *Nature* 405 (6788), 837–846.
- Perkins, T.J., Jaeger, J., Reinitz, J., Glass, L., 2006. Reverse engineering the gap gene network of *Drosophila melanogaster*. *PLoS Comput. Biol.* 2 (5), e51.
- Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., d’Alché-Buc, F., 2003. Gene networks inference using dynamic Bayesian networks. *Bioinformatics* 19 (Suppl. 2), ii138–ii148.
- Poliisetty, P.K., Voit, E.O., Gatzke, E.P., 2006. Identification of metabolic system parameters using global optimization methods. *Theor. Biol. Med. Model.* 3, 4.
- Quackenbush, J., 2002. Microarray data normalization and transformation. *Nat. Genet.* 32 (Suppl.), 496–501.
- Radke, D., Möller, U., 2004. Quantitative evaluation of established clustering methods for gene expression data. *Lect. Notes Comput. Sci.* 3337, 399–408.
- Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D.L., Falciani, F., 2004. Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics* 20 (9), 1361–1372.
- Rao, A., Hero III, A.O., States, D.J., Engel, J.D., 2007. Using directed information to build biologically relevant influence networks. *Comput. Syst. Bioinformatics Conf.* 6, 145–156.
- Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., Young, R.A., 2000. Genome-wide location and function of DNA binding proteins. *Science* 290 (5500), 2306–2309.
- Rice, J.J., Tu, Y., Stolovitzky, G., 2005. Reconstructing biological networks using conditional correlation analysis. *Bioinformatics* 21 (6), 765–773.
- Rodriguez-Fernandez, M., Mendes, P., Banga, J.R., 2006. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* 83 (2–3), 248–265.
- Rung, J., Schlitt, T., Brazma, A., Freivalds, K., Vilo, J., 2002. Building and analysing genome-wide gene disruption networks. *Bioinformatics* 18 (Suppl. 2), S202–S210.
- Sakamoto, E., Iba, H., 2001. Inferring a system of differential equations for a gene regulatory network by using genetic programming. In: Proceedings of the IEEE Congress on Evolutionary Computation. IEEE Press, pp. 720–726.
- Savageau, M.A., 1970. Biochemical Systems Analysis. Addison-Wesley, Reading 1970.
- Segal, E., Shapira, M., Regev, A., Pe’er, D., Botstein, D., Koller, D., Friedman, N., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34 (2), 166–176.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504.
- Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U., 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31 (1), 64–68.
- Soranzo, N., Bianconi, G., Altarini, C., 2007. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics* 23 (13), 1640–1647.
- Spieß, C., Streichert, F., Speer, N., Zell, A., 2005. Inferring regulatory systems with noisy pathway information. In: Proceeding of the German Conference on Bioinformatics—GCB 2005, Hamburg, Germany, pp. 193–203.
- Spieß, C., Hassis, N., Streichert, F., 2006. Comparing mathematical models on the problem of network inference. In: Proceeding of the 8th Annual Conference on Genetic and evolutionary computation (GECCO 2006), Washington, USA, pp. 279–285.
- Steuer, R., Kurths, J., Daub, C.O., Weise, J., Selbig, J., 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 18 (Suppl. 2), S231–S240.
- Stolovitzky, G., Monroe, D., Califano, A., 2007. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann. NY Acad. Sci.* 1115, 1–22.
- Stuart, J.M., Segal, E., Koller, D., Kim, S.K., 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302 (5643), 249–255.
- Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., Miyano, S., 2003. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics* 19 (Suppl. 2), ii227–ii236.
- Tegnér, J., Yeung, M.K.S., Hasty, J., Collins, J.J., 2003. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. U.S.A.* 100 (10), 5944–5949.
- Thomas, R., 1973. Boolean formalization of genetic control circuits. *J. Theor. Biol.* 42 (3), 563–585.
- Toepfer, S., Guthke, R., Driesch, D., Woetzel, D., Pfaff, M., 2007. The NetGenerator algorithm: reconstruction of gene regulatory networks. *Lect. Notes Bioinform.* 4366, 119–130.
- Van Berlo, R.J.P., van Someren, E.P., Reinders, M.J.T., 2003. Studying the conditions for learning dynamic Bayesian networks to discover genetic regulatory networks. *Simul.: Trans. Soc. Model. Simul. Int.* 79 (12), 689–702.
- Van Riel, N.A.W., 2006. Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief. Bioinform.*, 364–374.
- van Someren, E.P., Wessels, L., Reinders, M., 2000. Linear modeling of genetic networks from experimental data. In: Eight International Conference on Intelligent Systems for Molecular Biology, La Jolla, CA, USA, pp. 355–366.
- van Someren, E.P., Wessels, L.F., Backer, E., Reinders, M.J., 2001. Searching for limited connectivity in genetic network models. In: Proceeding of the 2nd International Conference on Systems Biology, Pasadena, California, pp. 222–230.
- van Someren, E.P., Wessels, L.F., Backer, E., Reinders, M.J., 2002a. Genetic network modeling. *Pharmacogenomics* 3, 507–525.
- van Someren, E.P., Wessels, L., Backer, E., 2002b. Regularization and noise injection for improving genetic network models. In: Computational and Statistical Approaches to Genomics. World Scientific Publishing Co, pp. 211–226.
- van Someren, E.P., Vaes, B.L.T., Steegenga, W.T., Sijbers, A.M., Dechering, K.J., Reinders, M.J.T., 2006. Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics* 22 (4), 477–484.
- Vilela, M., Chou, I.C., Vinga, S., Vasconcelos, A.T., Voit, E.O., Almeida, J.S., 2008. Parameter optimization in S-system models. *BMC Syst. Biol.* 16 (2), 35.
- Voit, E.O., 2000. Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists. Cambridge University Press, Cambridge, New York.
- Voit, E.O., 2008. Modelling metabolic networks using power-laws and S-systems. *Essays Biochem.* 45, 29–40.
- Wahde, M., Hertz, J., 2000. Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems* 55, 129–136.
- Wang, Y., Joshi, T., Zhang, X.S., Xu, D., Chen, L., 2006. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics* 22 (19), 2413–2420.
- Wasserman, W.W., Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5, 276–287.

- Weaver, D., Workman, C., Stormo, G., 1999. Modeling regulatory networks with weight matrices. In: Proceeding of the Pacific Symposium on Biocomputing, pp. 112–123.
- Werhli, A.V., Husmeier, D., 2007. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.*, 6:Article 15.
- Wessels, L.F.A., van Someren, E.P., Reinders, M.J.T., 2001. A Comparison of Genetic Network Models. Proceedings of the Pacific Symposium on Biocomputing, pp. 508–519.
- Westra, R., 2008. International Workshop on Gene Regulatory Network Inference, Jena, Personal Communication.
- Yeung, M.K.S., Tegnér, J., Collins, J.J., 2002. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. U.S.A.* 99 (9), 6163–6168.
- Yong-A-Poi, J., 2008. Adaptive least Absolute Regression Network Analysis Improves Genetic Network Reconstruction by Employing Prior Knowledge (Master Thesis). Delft University of Technology, The Netherlands.
- Zak, D.E., Doyle, F.J., Gonye, G.E., Schwaber, J.S., 2001. Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. In: Proceedings of the Second International Conference on Systems Biology, pp. 231–238.

## 4. Manuscript II

---

### **Integrative modeling of transcriptional regulation in response to antirheumatic therapy**

Michael Hecker, Robert Hermann Goertsches, Robby Engelmann,  
Hans-Jürgen Thiesen, and Reinhard Guthke

*BMC Bioinformatics* 2009, 10:262.

---

Methodology article

Open Access

**Integrative modeling of transcriptional regulation in response to antirheumatic therapy**Michael Hecker\*<sup>1</sup>, Robert Hermann Goertsches<sup>2</sup>, Robby Engelmann<sup>2</sup>, Hans-Juergen Thiesen<sup>2</sup> and Reinhard Guthke<sup>1</sup>

Address: <sup>1</sup>Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knoell-Institute, Beutenbergstr. 11a, D-07745 Jena, Germany and <sup>2</sup>University of Rostock, Institute of Immunology, Schillingallee 70, D-18055 Rostock, Germany

Email: Michael Hecker\* - michael.hecker@hki-jena.de; Robert Hermann Goertsches - robert.goertsches@med.uni-rostock.de; Robby Engelmann - roddy.engelmann@med.uni-rostock.de; Hans-Juergen Thiesen - hans-juergen.thiesen@med.uni-rostock.de; Reinhard Guthke - reinhard.guthke@hki-jena.de

\* Corresponding author

Published: 24 August 2009

Received: 10 February 2009

BMC Bioinformatics 2009, 10:262 doi:10.1186/1471-2105-10-262

Accepted: 24 August 2009

This article is available from: <http://www.biomedcentral.com/1471-2105/10/262>

© 2009 Hecker et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract**

**Background:** The investigation of gene regulatory networks is an important issue in molecular systems biology and significant progress has been made by combining different types of biological data. The purpose of this study was to characterize the transcriptional program induced by etanercept therapy in patients with rheumatoid arthritis (RA). Etanercept is known to reduce disease symptoms and progression in RA, but the underlying molecular mechanisms have not been fully elucidated.

**Results:** Using a DNA microarray dataset providing genome-wide expression profiles of 19 RA patients within the first week of therapy we identified significant transcriptional changes in 83 genes. Most of these genes are known to control the human body's immune response. A novel algorithm called TILAR was then applied to construct a linear network model of the genes' regulatory interactions. The inference method derives a model from the data based on the Least Angle Regression while incorporating DNA-binding site information. As a result we obtained a scale-free network that exhibits a self-regulating and highly parallel architecture, and reflects the pleiotropic immunological role of the therapeutic target TNF-alpha. Moreover, we could show that our integrative modeling strategy performs much better than algorithms using gene expression data alone.

**Conclusion:** We present TILAR, a method to deduce gene regulatory interactions from gene expression data by integrating information on transcription factor binding sites. The inferred network uncovers gene regulatory effects in response to etanercept and thus provides useful hypotheses about the drug's mechanisms of action.

**Background**

The molecular interactions within a biological system give rise to the function and behavior of that system. In systems biology, one aims to formulate the complex interac-

tions of biological processes by mathematical models. A major focus of the field is the uncovering of the dynamic and intertwined nature of gene regulation.

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

Gene expression is mainly regulated at the level of mRNA transcription by proteins called transcription factors (TFs), that specifically bind the DNA at the regulatory region of their target genes. A number of collections of experimentally defined TF binding sites (TFBS) have been assembled. The most commonly used is the Transfac database, which catalogs eukaryotic TFs and their known binding sites [1]. The expression level of a gene usually depends on the occupancy states of multiple TFBS. However, gene regulation is much more complex and includes different layers of post-transcriptional control. The entirety of gene regulatory processes constitutes a network of genes, regulators, and the regulatory connections between them – namely a gene regulatory network (GRN). In the past, various modeling approaches have been proposed to (partially) reconstruct GRNs from experimental data on the basis of different mathematical concepts and learning strategies, and distinct degrees of abstraction [2-4]. A graph is always the basic modeling scheme for a GRN, with nodes symbolizing regulatory elements (e.g. genes and proteins) and edges representing (activatory and inhibitory) relationships between them. Common mathematical formalisms of such a graph are Boolean networks, Bayesian networks, association networks and systems of equations. Boolean networks assume that genes are simply on or off, and apply Boolean logic to model dynamic regulatory effects. In contrast, Bayesian networks model gene expression by random variables and quantify interactions by conditional probabilities. Interactions in association networks are typically undirected and derived by analyzing pairs of genes for co-expression e.g. using mutual information as a similarity measure. Systems of equations describe each gene's expression level as a function of the levels of its putative predictors. For specific types of functions they could draw on well developed statistical techniques to efficiently fit their model parameters. However, GRN inference is always a challenging task because of incomplete knowledge of the molecules involved, the combinatorial nature of the problem and the fact, that often available data are limited and inaccurate. Microarray gene expression data are typically used to derive rather phenomenological GRN models of how the expression level of a gene is influenced by the expression level of other genes, i.e. the model also includes indirect regulatory mechanisms. Obviously, the incorporation of other types of data in addition to gene expression data (e.g. gene functional annotations, genome sequence data, protein-protein and protein-DNA interaction data) as well as the integration of prior biological knowledge (e.g. from scientific literature) supports the inference process. Moreover, it is necessary to utilize biological plausible assumptions considering the network topology (e.g. structural sparseness). The integration of diverse types of biological information and modeling constraints allows for more accurate GRN models and is a

current challenge in network reconstruction. Bayesian networks and systems of linear equations have been most studied for such combined analyses [3-5].

Organizing biological data in network models may help understanding complex diseases such as human autoimmune diseases [6]. Many studies implicate hundreds of genes in the pathogenesis of autoimmune diseases, but we still lack a comprehensive conception of how autoimmunity arises. Understanding structure and dynamics of molecular networks is critical to unravel such complex diseases. Network analyses may not only support the investigation of autoimmune diseases but also the optimization of their treatment. Here, we focus on rheumatoid arthritis (RA), which is a multifactorial polygenic disease and might be termed a systems biology disease. RA is a chronic inflammatory disorder primarily afflicting the synovial joints, and autoimmunity plays a pivotal role in its chronicity and progression. The disease is characterized by autoreactive behavior of immune cells and the induction of enzymes which lead to the destruction of cartilage and bone [7]. The inflammatory processes are triggered by cytokines and other immune system-related genes that form a complex network of intra- and intercellular molecular interactions. A number of cytokine proteins play a critical role as mediators of immune regulation. In RA, the two cytokines TNF-alpha and IL-1 are considered master regulators that act in a complementary and synergistic manner [8,9]. By blocking TNF-alpha, etanercept intervenes this molecular network and thus is thought to rebalance the immune system's dysregulation [10-12]. Etanercept therapy in RA patients has been proven to slow disease progression, but the precise molecular mechanisms remained unclear. To investigate the therapeutic effects on transcriptional regulation, GRN inference techniques can be applied. This could lead to a better understanding of the modes of action of etanercept as well as the pathogenesis underlying RA. We may also understand why the drug fails to control the disease in about 30% of the patients (non-responders).

We studied a group of 19 patients suffering from RA for which DNA microarrays were used to obtain genome-wide transcriptional profiles within the first week of etanercept administration [13]. A set of etanercept responsive genes was attained. The majority of these genes are known to control the body's immune response. Several TFBS were identified as overrepresented in the genes' regulatory regions and we used the corresponding information on TF-gene interactions as a template for modeling the underlying GRN. A system of linear equations was chosen to mathematically describe the regulatory effects between the genes and TFs (i.e. the network nodes). We used a hybrid of the Least Angle Regression (LARS) and the Ordinary Least Squares regression (OLS) to find the

model structure and estimate the coefficients. In doing so, the modeling is constrained to include only a subset of the putative TF-gene interactions. That way, our approach considers that genes usually regulate other genes indirectly through the activity of one or more TFs, which makes the model straightforward to interpret in terms of true molecular interactions. The resulting GRN was further analyzed using e.g. gene ontology (GO) and clinical information (figure 1). We found that our integrative modeling strategy, namely the TFBS-integrating LARS (TILAR), is able to reconstruct GRNs more reliably than other established methods. This is one of the first studies that utilizes network analysis to investigate transcriptional regulation in response to a therapeutic drug in humans [14].

## Results and discussion

### Effects of etanercept therapy on gene expression

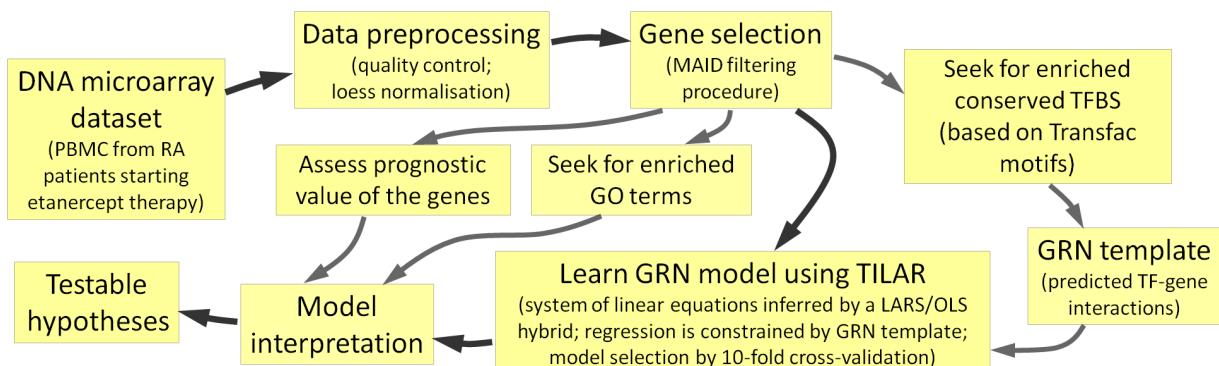
We used the Affymetrix microarray dataset from Koczan *et al.* [13] which provided expression levels of peripheral blood mononuclear cells (PBMC) measured in 19 patients suffering from RA. For each patient, blood samples were taken before treatment (baseline) as well as 72 (day 3) and 144 hours (day 6) after start of immunotherapy by etanercept. Clinical response was assessed over 3 months and revealed 7 patients with persistent disease activity (non-responders).

We analyzed the DNA microarray data in respect to common gene expression changes observed in the whole group of patients after therapy onset. First of all, we pre-processed the data to correct for systematic effects. More importantly, signal intensities were calculated by applying a custom chip definition file by Ferrari *et al.* that is composed of custom-probesets including only probes matching a single gene [15]. As such, a one-to-one correspondence between genes and custom-probesets is

preserved, which deeply improves gene-centered analysis of human Affymetrix data [16]. Finally, the data pre-processing yields expression levels of 11,174 different genes for each of the 55 microarrays in the dataset (for details see the methods section).

Afterwards, we identified a set of genes significantly up- or down-regulated in response to etanercept. It is important to note that the filtering of genes is a crucial step in GRN inference as there is a tight relationship between model complexity (i.e. network size and level of detail of the model), the amount of data required for inference and the quality of the results. On the one hand, a small and detailed network model might better fit the given data, but only a sufficiently large model can capture the fundamental properties that constitute a GRN including scale-freeness, redundancy and self-regulation. In this study, we utilized a *t*-statistic in conjunction with an MA-plot-based signal intensity-dependent fold-change criterion (referred to as MAID filtering) to select genes with expression changes in the first week of therapy (see methods). Through this filtering we identified 37 genes as differentially expressed at day 3 versus baseline, and 57 genes at day 6. Altogether, 48 genes were found down-regulated and 35 genes up-regulated, comprising a set of 83 genes in total (additional files 1 and 2).

We searched for overrepresented terms of the GO biological process ontology in the list of 83 selected genes and found that most of the genes are known to control the body's immune response (additional file 3, see methods). Remarkably, genes of the I-kappaB kinase/NF-kappaB cascade (GO:0007249) are enriched in the gene set and represented by 5 genes (NFKBIA, TNFRSF1A, TLR8, NOD2, HMOX1). NF-kappaB is a key factor in the transcription of many inflammatory genes and has been implicated in the pathological processes of RA. The NF-kappaB cascade is



**Figure 1**  
**Workflow used to study gene regulatory effects in response to etanercept therapy.** A network model of transcriptional regulation is inferred by integrating transcription factor binding site information.

mainly activated by the proinflammatory cytokines IL-1 and TNF-alpha. As was shown, TNF-blocking agents such as etanercept prevent TNF-alpha from binding to its receptors, induction of signal transduction cascades and activation of TFs including NF-kappaB [10]. Here, we found NFKBIA, that inactivates NF-kappaB by trapping it in the cytoplasm, up-regulated early after therapy initiation. On the other hand, genes known to activate the NF-kappaB protein, e.g. NOD2 and TNFRSF1A (a TNF receptor), were down-regulated after therapy onset. Thus, the result of our filtering indicates the expected suppression of NF-kappaB activity by etanercept. In addition, we found evidence for a modulation of B cell mediated immunity. The corresponding GO category (GO:0019724) comprises the genes C1QB, CLU and TLR8 whose mean expression was significantly lower at day 3 and 6 compared to baseline, respectively. Interestingly, TLR8 signaling is linked to the control of CD4+ regulatory T (Treg) cells. Treg cells actively suppress host immune responses and, as a consequence, play an important role in preventing autoimmunity [17]. TLR8 is thought to initiate immune processes by reversing the suppressive function of Treg cells [18]. Its down-regulation by etanercept might be an important factor to control the disease.

Genes responsive to etanercept administration are probably under control of certain TFs, whose activities are (maybe indirectly) affected by this drug. Therefore, we analyzed the regulatory regions around the respective transcription start sites (TSS) of these genes for occurrence of overrepresented TFBS (see methods). Identifying TFBS, particularly in higher eukaryotic genomes, is an enormous challenge and cross-species sequence conservation is often used as an effective filter to improve the predictions. We found evolutionarily conserved binding sites enriched for 12 TFs (represented by 19 Transfac binding profiles).

These 12 TFs connect 52 out of the 83 genes through 96 TF-gene interactions, whereas each TF is linked to at least 4 genes (table 1). The list of TFs includes C/EBP-beta, which is an important transcriptional activator in the regulation of genes involved in immune and inflammatory responses, including the cytokines IL-6, IL-8 and TNF-alpha [19]. Binding sites for the TATA binding protein (TBP) were detected in 14 genes. TBP binds DNA at the TATA-element, and as a subunit of the TFIID complex coordinates the initiation of transcription by RNA polymerase. Although TBP is always involved, its TATA-binding activity is dispensable for the positioning of the RNA polymerase. In fact, approximately 76% of human core promoters lack TATA-like elements [20]. However, in the set of 83 genes, those genes having the TATA box were overrepresented. The two TFs ZIC1 and ZIC3 were considered as one TF entity, as they have highly similar DNA binding properties. None of the 12 TFs showed significant transcriptional changes in the data. Nevertheless, the information on predicted TF-gene interactions can be used as a GRN template during inference. Before describing how this is done by TILAR we will outline the general principles of the modeling approach.

#### Linear network modeling

We chose a system of equations to model the regulatory interactions among the genes affected by etanercept therapy. The concept of modeling gene regulation by a system of equations is to approximate gene expression levels as a function of the expression of other genes and environmental factors. Modeling GRNs by systems of equations has several benefits as they can describe regulatory effects in a flexible, quantitative, directed manner, and take into account that gene regulators act in combination. With systems of equations one can easily model positive and negative feedback loops, and describe even non-linear and

**Table 1: Evolutionarily conserved binding sites were found to be enriched for 12 TFs.**

TF Name	Transfac ID	Official Full Name	P-value	Expected Count	Count
TBP, TFIID	V\$TBP_01, V\$TATA_C, V\$TATA_01	TATA box binding protein	0.0042	6.41	14
C/EBPbeta	V\$CEPB_01, V\$CEPB_02	CCAAT/enhancer binding protein beta	0.0112	5.03	11
Zic1, Zic3	V\$ZIC1_01, V\$ZIC3_01	Zic family member 1/3	0.0183	6.13	12
AP-2rep	V\$AP2REP_01	Kruppel-like factor 12	0.0264	1.68	5
HNF-1, HNF-1A	V\$HNF1_01, V\$HNF1_C	HNF1 homeobox A	0.0274	2.30	6
Lmo2	V\$LMO2COM_01, V\$LMO2COM_02	LIM domain only 2	0.0352	5.98	11
SRY	V\$SRY_02	sex determining region Y	0.0374	1.85	5
ATF-2	V\$CREBP1_01	activating transcription factor 2	0.0408	1.30	4
Cart-1	V\$CART1_01	ALX homeobox 1	0.0415	1.30	4
COMPI	V\$COMPI_01	cooperates with myogenic proteins 1	0.0422	3.23	7
Hlf	V\$HLF_01	hepatic leukemia factor	0.0470	1.97	5
NF-1, NF-1/L	V\$MYOGNFI_01, V\$NFI_Q6	nuclear factor 1	0.0492	7.10	12

$\Sigma = 96$

The column "Count" denotes the number of genes that possess a TFBS for the respective TF. All in all, 96 TF-gene interactions were predicted (GRN template).

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

dynamic phenomena of biological systems. However, as more complex models require higher amounts of accurate data to learn their parameters reliably, researchers often utilize systems of linear equations (linear models). Linear models have been successfully employed in many applications, e.g. to reconstruct GRNs relevant for development of the central nervous system in rats [21], osteoblast differentiation in mice [22,23], galactose regulation in yeast [24] and immune response of human blood cells to bacterial infection [25]. Linear models assume that gene regulatory effects are limited to be linear and additive and a simple one can be written as:

$$\hat{x}_i = \sum_{j=1, j \neq i}^N w_{ij} x_j, \quad (1)$$

where vector  $x_i$  contains the  $M$  expression levels measured for gene  $i$ ,  $N$  is the number of genes in the network, and the weights  $w_{ij}$  define relationships between the genes. When inferring a linear model we need to estimate the weights  $w_{ij}$  (i.e. the model parameters) from the data. The weights specify the existence of regulatory relationships between genes, their nature (activation or inhibition) and relative strength. If  $w_{ij} > 0$  gene  $x_j$  activates gene  $x_i$ , if  $w_{ij} < 0$   $x_j$  inhibits  $x_i$ , and  $w_{ij} = 0$  implies that  $x_i$  is not under control of  $x_j$ . This simplicity makes linear models easy to interpret, even if the encoded relationships have a wide range of meanings: edges in the network might represent direct physical interactions (e.g. when a gene encodes a TF regulating another gene) or rather conceptual interactions (e.g. when the expression levels of two genes merely correlate).

Linear models can also be used to describe the dynamics of the network. In this case, the model is a system of linear difference equations that approximates the change of gene expression in time. However, this approach is inappropriate for our application as the time-series in the microarray dataset consist of only 3 time-points and the time between two subsequent measurements is rather long (3 days). Nevertheless, the modeling strategy illustrated here can be easily adapted for the inference of dynamic models.

To fit the (static) linear model to the data, equation (1) can be written in matrix form as follows:

$$\begin{aligned} \hat{y}^i &= X_{M \times (N-1)}^i \cdot \beta^i \text{ for } i = 1 \dots N, \text{ where} \\ y^i &= x_i, X_{M \times (N-1)}^i = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N), \beta^i = (w_{i1}, \dots, w_{i,i-1}, w_{i,i+1}, \dots, w_{i,N})^T. \end{aligned} \quad (2)$$

These  $N$  systems can be coupled as:

$$\hat{y} = X_{MN \times (N-1)N} \cdot \beta \text{ or in shorter form : } \hat{y} = X \cdot \beta. \quad (3)$$

Now, a GRN model can be inferred by estimating  $\beta$  (comprising all the model parameters in  $w$ ) from input matrix  $X$  (having  $M' = MN$  rows and  $N' = (N-1)N$  columns) and output vector  $y$  using OLS regression.

However, despite the fact that linear models are a strong simplification of the true GRN, equation (3) is already an underdetermined system of linear equations in our particular study as the number of genes in the network ( $N = 83$ ) is greater than the number of measurements ( $M = 55$ ). That means, infinitely many solutions exist. Therefore, biologically motivated constraints have to be included to tackle this problem. The most commonly used modeling constraint is the sparseness of GRNs. Sparseness reflects the fact that genes are regulated only by a limited number of regulators. The sparseness constraint minimizes the number of edges, i.e. reduces the effective number of model parameters. Sparse linear models can be reconstructed via the Lasso (Least absolute shrinkage and selection operator) method [26], which effectively performs simultaneous parameter estimation and variable selection. The Lasso is a version of OLS that constrains the sum of the absolute regression coefficients  $\beta$ :

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^{N'} \left( y_i - \sum_{j=1}^{M'} x_{ij} \cdot \beta_j \right)^2 \right\} \text{ subject to } \sum_{j=1}^{M'} |\beta_j| \leq s. \quad (4)$$

The Lasso penalizes model complexity by shrinking the coefficients  $\beta_j$  (and hence  $w_{ij}$ ) toward 0, more so for small values of  $s$ . A modification of the LARS algorithm implements the Lasso [27]. LARS builds up estimates for  $\beta$  in successive steps, each step adding one covariate to the model, so that gradually model parameters are set non-zero. In simple terms, LARS is a less greedy version of traditional forward selection methods. LARS and its variants are computationally efficient. The algorithm requires only the same order of magnitude of computational effort as OLS to calculate the full set of Lasso estimates (i.e. for all  $s \geq 0$ ).

The Lasso approach was first introduced to infer regulatory interactions by van Someren *et al.* [28] and has since been applied in several GRN studies [22,29]. However, even if the network connectivity is constrained, there is a limitation in inferring GRNs using gene expression data only. Hence, there is a need to incorporate different types of information during network reconstruction. Various data and information from biomedical literature and

databases can be utilized in combination with gene expression levels to increase model accuracy.

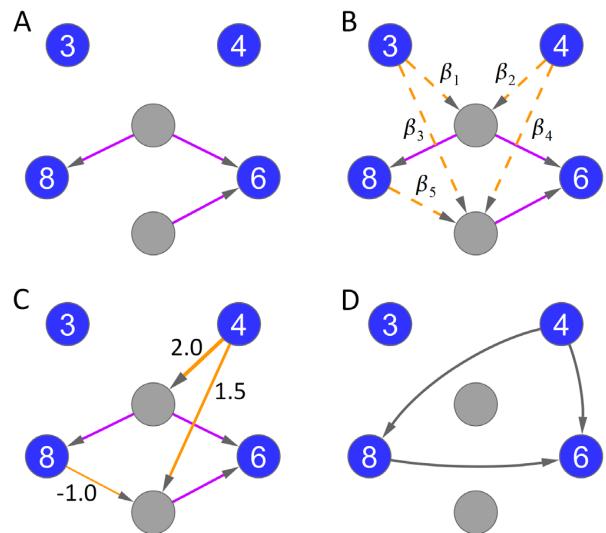
An integrative learning strategy usually consists of two steps. First, a template of the network is built, e.g. based on known TF-DNA interactions or molecular interactions automatically extracted from the literature by text mining. This template represents a supposition of the true network structure, that might be uncertain and incomplete. Second, an inference algorithm is applied that fits the model to the measured data while taking the template into account, trading off data-fit and template-fit. When inferring linear models, such template information can be included by adapting the Lasso method. This is possible by introducing additional weights  $\delta$  on the coefficients  $\beta$  of the constraint in equation (4):

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^{N'} \left( y_i - \sum_{j=1}^{M'} x_{ij} \cdot \beta_j \right)^2 \right\} \text{ subject to } \sum_{j=1}^{M'} \delta_j |\beta_j| \leq s. \quad (5)$$

A relatively low weight  $\delta_j$  provokes that the edge corresponding to  $\beta_j$  is preferred to be in the final model. Hence, the modeler is able to incorporate partial prior knowledge by setting the weights  $\delta$  appropriately. Recently, this concept was applied to integrate human microarray data with regulatory relationships obtained by literature mining by defining each  $\delta_j$  as a constant [23] and as weight function of  $\beta_j$  [13], respectively.

#### TILAR – a TFBS-integrating linear modeling approach

Here, we propose TILAR – a TFBS-integrating inference technique that differs from the adaptive Lasso approach, and employs TF binding information as prior knowledge. As we will show, it is even possible to combine the adaptive Lasso and TILAR. According to our modeling scheme, we distinguish two types of network nodes: genes (that were selected for inferring regulatory relationships between them) and TFs (for which respective binding sites are overrepresented in the gene set). Expression levels of the gene set (that possibly includes genes encoding TFs) are required for the modeling. The algorithm then aims to assign (directed) TF-gene and gene-TF interactions (network edges). A TF-gene interaction represents a physical interaction, i.e. a TF binds the region that encompasses the TSS of a certain gene and thus regulates its transcription. In contrast, gene-TF interactions can have different meanings: the gene itself might encode a transcriptional regulator of the TF, or the gene product controls the activity of the TF at the proteomic level, or the gene triggers signaling cascades that affect the TF, etc. Using both types of interactions, the model reflects that genes regulate other genes indirectly through a combination of TFs (figure 2). As a reminder, the TFBS overrepresentation analysis



**Figure 2**  
**Illustration of our proposed modeling conception.**  
Here, we aim to reconstruct a gene regulatory network consisting of 4 genes (dark blue) and 2 transcription factors (light gray). For simplicity, we assume that only one gene expression measurement was performed. The expression level of each gene is given in the gene nodes. (A) The GRN template: In this example, two genes possess at least one TF binding site in their regulatory region as indicated by 3 TF-gene interactions (purple). (B) In that case, there are 5 possible gene-TF interactions (i.e. model parameters  $\beta$ ) in the network (dashed, orange). If available, we might consider prior knowledge on gene-TF interactions during inference (adaptive TILAR). (C) A possible inference result including 3 gene-TF interactions (solid, orange). Here, the model perfectly fits the data (e.g. "8" = 2.0 · "4") with two nominal model parameters set to zero. (D) We can use the inferred model to derive gene-gene relationships from the edges between genes and TFs (gray). The benchmarking was conducted on such gene-gene interactions.

revealed 96 putative TF-gene interactions. Now, the idea is to use this information as a GRN template by constraining the modeling to include only a subset of these TF-gene interactions. As the inference method not necessarily uses all the given TF-gene interactions, we consider the fact that they are computationally predicted and therefore not all of them might refer to biologically functional binding sites. In practice, the algorithm starts with the entire set of TF-gene interactions and then iteratively removes avoidable interactions through a backward stepwise selection procedure (see methods). The information on (the current set of included) TF-gene interactions is written in matrix  $B$ , which is defined as:

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

$$b_{kj} = \begin{cases} 1, & \text{if gene } j \text{ possesses a binding site for TF } k \\ 0, & \text{else.} \end{cases} \quad (6)$$

In our particular study, 96 entries in  $B$  were set to 1 in the first iteration. TILAR then assigns the parameters in the model to gene-TF interactions, as follows:

$$\hat{x}_i = \sum_{k=1}^F \sum_{j=1}^N (1 - b_{kj}) w_{kj} x_j b_{ki}, \quad (7)$$

where  $F$  is the number of TFs. For modeling the transcriptional regulation in response to etanercept we have  $N = 83$  genes, that showed significantly changed expression levels after therapy onset, and  $F = 12$  TFs, whose binding sites are overrepresented in the regulatory regions of the selected genes. In the model, each gene can exhibit a regulatory effect on each TF, except those TFs that hold a TF-gene interaction to this gene (this restriction is dispensable when inferring a dynamic model). If  $w_{kj} = 0$ , there is no gene-TF interaction between gene  $j$  and TF  $k$ . Otherwise, gene  $j$  controls the activity of TF  $k$  and thus regulates all the genes that possess a TFBS for TF  $k$ . In this case, the expression levels of gene  $j$  explain the expression of the genes regulated by TF  $k$ . Again, to infer the GRN model, we need to estimate the model parameters  $w_{kj}$  from the gene expression data while constraining the model to be sparse. Similar to equation (3), we can couple the subsystems, as follows:

$$\hat{y} = X_{MN_r \times (FN - \#B)} \cdot \beta, \quad (8)$$

where  $N_r$  is the number of genes possessing at least one overrepresented TFBS, and  $\#B$  is the number of TF-gene interactions considered at the current iteration. The coefficients  $\beta$  of this equation now correspond to gene-TF interactions in the model. Finally, a sparse solution to equation (8) can be found using the Lasso according to equation (4).

The TILAR modeling approach proposed here is advantageous for several reasons. First, TF expression levels are not required, since the activity of TFs is modeled implicitly (like a hidden node). This is beneficial, as mRNA levels of TFs are often low and do not necessarily correlate with TF activity. In fact, TF proteins often need to be activated by phosphorylation. Second, the nominal number of model parameters  $w$  in equation (7) is generally lower than in equation (1). Therefore, our method tackles the problem of having too many parameters in comparison to limited amounts of experimental data. In our particular application, equation (8) is an overdetermined system of linear equations (as  $M \cdot N_r = 55 \cdot 52 = 2860 > F \cdot N - \#B =$

$12 \cdot 83 - \#B = 996 - \#B$ ,  $\#B \leq 96$ ), i.e. we are able to infer a complex network of 83 genes (and 12 TFs) without being in conflict with the data requirements. Third, by using TF binding predictions as prior knowledge we can reconstruct GRNs more reliably. Besides, the inferred models are relatively easy to interpret. Finally, the integration of TFBS information is accomplished by simply specifying the regression equation (i.e. input matrix  $X$  and output vector  $y$ ) adequately. Therefore, we can combine the TILAR approach with the adaptive LARS, i.e. solve equation (8) according to equation (5) if prior knowledge on gene-TF interactions is available (adaptive TILAR).

#### Modeling the gene regulatory response to etanercept

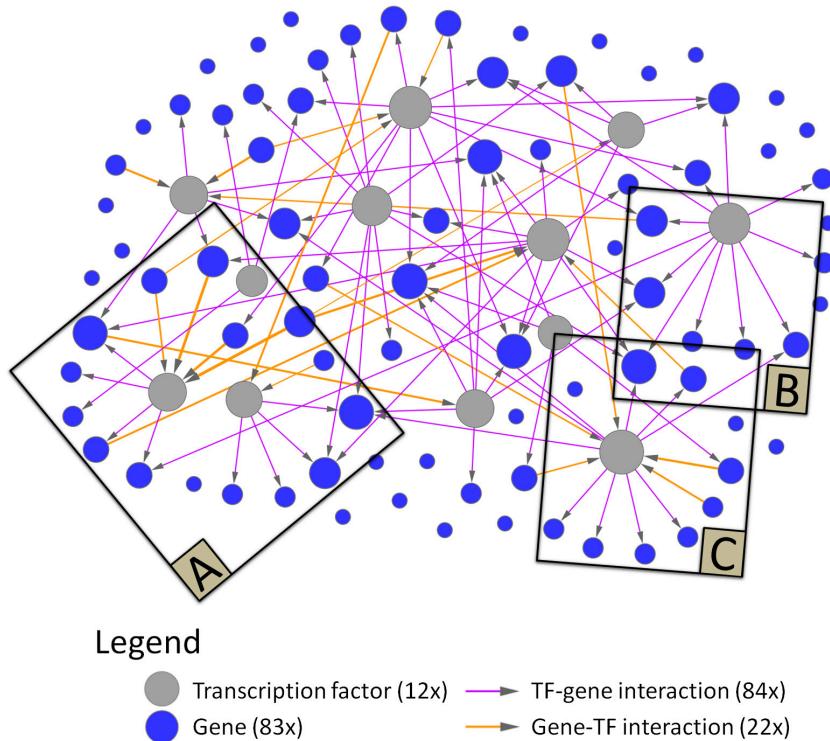
To examine the early transcriptional effects of etanercept we applied the TILAR algorithm to construct a GRN model on the basis of gene expression data and knowledge on TF-gene interactions obtained by TFBS analysis. For this purpose, the GRN inference problem was formulated according to equation (8). The essential part of our modeling approach is the LARS algorithm that is used to obtain all possible Lasso solutions for this linear regression equation.

TILAR iteratively applies LARS in a backward stepwise selection procedure in order to refuse TF-gene interactions that do not fit the data well (see methods). Hence, the learning strategy takes into account that the prediction of TFBS might be error-prone. In this study, 12 out of 96 predicted TF-gene interactions were discarded. These 12 interactions may result from false positive TFBS predictions, or the magnitude of the TF-gene interactions was not enough for being confirmed based on the gene expression levels.

After we identified the subset of 84 TF-gene interactions, we used LARS to define which gene-TF interactions have to be included at different degrees of network connectivity. That means we used LARS only for variable selection, but the actual coefficients were estimated by OLS (see methods). This LARS/OLS hybrid technique usually achieves sparser estimates and more accurate predictions, and thus outperforms the ordinary Lasso [27,30]. Finally, we selected the most parsimonious estimate with low 10-fold cross-validation error (additional file 4). In this way, the method avoids overfitting to the data and consequently yields a sparse GRN model. The final model consists of 22 inferred gene-TF interactions and 84 TF-gene interactions, and was visualized using Cytoscape 2.6.0 (figure 3, additional file 5).

#### Model interpretation

Systems biological models need to be interpretable in order to be useful. In general, the modeling goals of accurate prediction and interpretation are contradictory since



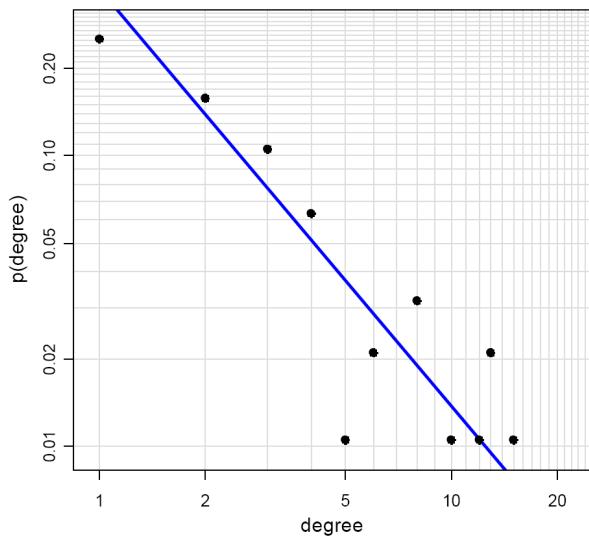
**Figure 3**  
**Reconstructed gene regulatory network of genes up- or down-regulated during first week of therapy.** The TILAR algorithm used gene expression data and transcription factor binding predictions to infer a network of 84 TF-gene and 22 gene-TF interactions. The size of the nodes corresponds to their degree of connectivity. Three parts of the network model are shown in detail in figure 5. The full model is available as a Cytoscape session file of (additional file 5).

interpretable models should be simple, but more accurate models might be quite complex. The model reconstructed here seems to satisfy both requirements. On the one hand, the network model is fairly complex as it consists of 95 nodes and 106 edges while using 22 model parameters  $w_{ij}$  (specifying the strength of the gene-TF interactions). Yet the model is readily interpretable due to the intuitive linear modeling scheme.

Apparently, the inferred model is sparse, i.e. each network node is under control of only few regulators. The maximum in-degree in the GRN is 5 (on average 1.12). Nevertheless, some nodes (named hubs) are highly connected in the network, e.g. TBP which has an out-degree of 12. A further characteristic and biologically meaningful property of the network is its scale-free structure. Scale-freeness denotes the phenomenon that the degree distribution in biological networks often follows a power law, i.e. the fraction  $P(k)$  of nodes in the network having  $k$  connections goes as  $P(k) \sim k^{-\gamma}$ , where  $\gamma$  is a constant. This means

that in scale-free networks most of the nodes are lowly connected, while a few are relatively highly connected. Scale-freeness indicates a network's decentralization and structural stability, and in consequence its robustness against random fluctuations [31]. The scale-free design of GRNs is well studied in literature [32,33], and the GRN reconstructed here is scale-free with  $\gamma = 2.22$  (as calculated according to Clauset *et al.* [34], see figure 4).

A closer look at the interactions in the network revealed gene sets co-regulated by a common TF. For example, 6 TF-gene interactions were assigned to the transcriptional activator HNF-1 in the GRN template (table 1). Two of them were not considered in the final model as they were eliminated during backward stepwise selection. However, the 4 remaining genes that are predicted to be under control of HNF-1 (AQP9, TCN2, CREB5, C4orf18) are all down-regulated in the patients during first week of therapy (figure 5A). AQP9 is assumed to have some role in immunological response [35]. Hence, we can hypothesize



**Figure 4**  
**Node degree distribution in log-log scale.** The network is scale-free, while transcription factors are more connected than genes. The orthogonal linear regression line is shown in blue.

that the activity of HNF-1 is lowered under etanercept therapy, which has (barely explored) effects on specific immune processes.

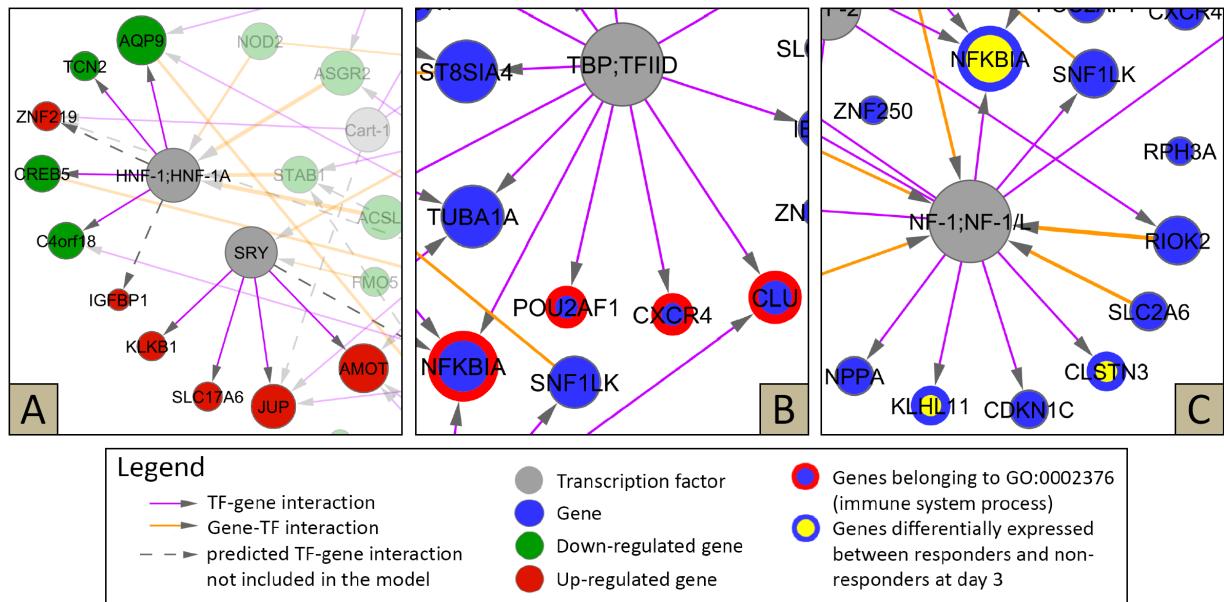
We also found that the network model highlights TFs that regulate functionally related genes (as annotated by GO). For instance, the model reveals TF-gene interactions of the transcription initiation factor TBP and the genes NFKBIA, POU2AF1, CXCR4 and CLU (figure 5B). These 4 genes not only share the TATA binding site in their regulatory region, but also belong to the same functional category (immune system process, GO:0002376). Nevertheless, they play different roles in inflammatory control. NFKBIA inhibits the activity of the NF-kappaB complex, which controls many genes involved in inflammation and is chronically active in RA [10]. Interestingly, the data show significantly elevated expression levels of NFKBIA in response to etanercept. POU2AF1 is a B cell-specific transcriptional co-activator that is known to stimulate immunoglobulin promoter activity [36], and CXCR4 is a chemokine (C-X-C motif) receptor that guides lymphocyte migration [37]. These findings suggest that the therapy by etanercept modulates the maladjusted immune system at multiple levels.

Other important features of a GRN are feedback and redundancy mechanisms. Regulatory feedback loops can be positive (i.e. reinforcing) or negative (i.e. self-balancing). Redundant links in the GRN allow genes to maintain

their connection to other genes even if some genes are malfunctioning. Redundancy and self-control provide flexibility and adaptability to environmental changes, i.e. robustness against noise and failures [31]. An exemplary (positive) feedback loop in the inferred GRN model is the regulatory chain "CREB5 → C/EBP-beta → ASGR2 → HNF-1 → CREB5". Notably, C/EBP-beta encodes a TF that is important in the regulation of immune genes and has been shown to bind the regulatory regions of several cytokine and acute-phase genes. In RA, elevated levels of acute-phase proteins have been associated with progressive joint damage [38]. The feedback loop is finally formed by the two gene nodes CREB5 (which encodes a TF as well) and ASGR2. Both genes were down-regulated after therapy onset. Therefore, we assume that etanercept lowers the activity of C/EBP-beta while affecting a regulatory feedback mechanism.

The GRN model also contains a (positive) feedforward loop composed of the two ways "NOD2 → HNF-1" and "NOD2 → Lmo2 → STAB1 → HNF-1". NOD2 is a regulator of NF-kappaB activity [39] and was found down-regulated on days 3 and 6. The model predicts a gene-TF interaction between NOD2 and HNF-1, while we presume a decreased activity for HNF-1 as described previously. Alternatively, NOD2 is linked to LMO2, which has a crucial role in hematopoietic development and is connected to STAB1 according to the model. In turn, STAB1, a receptor which is supposed to function in angiogenesis and lymphocyte homing [40], has a gene-TF interaction to HNF-1, thereby closing the feedforward loop. Ultimately, this demonstrates the cooperative action of genes in the network.

As mentioned before, out of the 19 RA patients in the analyzed dataset 7 did not respond clinically to etanercept. Considering the potential side effects and the high costs of the therapy, the identification of patients who will most likely respond would contribute to a more optimized treatment of RA. To identify predictors (biomarkers) of the therapeutic outcome one might seek for differences in the gene expression of responder and non-responder patients before or early in therapy. The work by Koczan *et al.* is focused on this particular issue [13]. Here, we also compared the transcriptional levels of both patient groups using a two-sample *t*-test. At day 3, four network genes were found to be differentially expressed at the significance level  $\alpha = 0.05$  (additional file 1). Three of these genes (NFKBIA, KLHL11, CLSTN3) were expressed lower in the responder group and are regulated by a common TF node (NF-1) in the GRN model (figure 5C). NF-1 (nuclear factor I) constitutes a family of DNA-binding proteins with similar binding specificity, that participate in both cell type-specific transcription and replication [41]. Our model suggests that NF-1 regulates genes that are possibly



**Figure 5**  
**Detail views of the network model shown in figure 3.** (A) The modeling strategy takes into account that target genes of a transcription factor are often co-expressed. For example, all the genes that are regulated by HNF-1 are down-regulated after therapy onset. Outer parts are shown with lower opacity. (B) A set of genes associated with the GO category "immune system process" is predicted to contain TATA-like elements in their regulatory regions. (C) Three genes were expressed lower in responders at day 3 and all of them are regulated by NF-1 according to the model.

relevant for the individual success of etanercept therapy, while the prognostic value of NFKBIA mRNA levels is already under discussion [13]. However, even if this hypothesis still has to be verified, the analysis clearly demonstrates the use of correlating clinical features with molecular network structures [6,42].

The GRN model provides many testable hypotheses and thus may be a starting point for new experiments. Those could aim to study the expression changes of specific genes in more detail, or to analyze their regulatory effects thereby validating parts of the inferred network. Take for instance the previously mentioned subnetwork of NF-1 and its target genes (figure 5C). The model suggests that members of the NF-1 family bind the promoters of 10 etanercept-responsive genes. This might be tested by electrophoretic mobility shift assays or chromatin immunoprecipitation techniques. One could also investigate the genes' transcriptional changes during therapy in a larger cohort of RA patients by generating expression profiles using real-time polymerase chain reaction. This would be particularly useful for the three genes that were found significantly lower expressed in the responder group. As a next step, levels of the respective proteins and protein isoforms could be quantified using western plots and enzyme-linked immunosorbent assays. For example, it

would be interesting to measure the amount of HNF-1 proteins as we postulate a lowered activity of this TF as a molecular therapeutic effect. In the model all target genes of HNF-1 are down-regulated in response to etanercept administration (figure 5A). Similarly, other parts of the network are worth to study, e.g. the inferred regulatory feedback loop including C/EBP-beta and CREB5. Transcript and protein levels of *in vitro* cultures of PBMC cells may also be analyzed in a time-dependent manner. This allows for controlled perturbation experiments such as siRNA mediated knock-down of NF-1 expression with or without the presence of etanercept. Last but not least, one could examine the cell type-specific expression of genes in the network. Recent studies point out a functional impairment of Treg cells in RA [17]. It would be attractive to further elucidate the altered immunosuppressive capacity of Treg cells, their role in the treatment of RA and the modulation of Treg cells by Toll-like receptors such as TLR8 that was down-regulated in the dataset.

#### Performance evaluation

To demonstrate the benefit of the TILAR modeling approach, we tried to evaluate how reliable the structure of the underlying GRN can be inferred. The assessment of the GRN inference performance is a challenging task, as evidently, true regulatory interactions are barely known

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

and curated datasets for benchmarking are missing, though there are attempts to remedy this shortcoming [43,44]. A further difficulty is that the knowledge used to validate a GRN model must be different from the knowledge integrated during modeling.

Here, we utilized gene-gene interaction information obtained by text mining for performance evaluation (see methods). By assessing the inference quality on literature-derived (undirected) gene-gene links, we were also able to compare our method with other inference techniques which do not incorporate prior knowledge. It is important to note that regulatory gene-gene interactions are implicitly defined in our network model by gene-TF and TF-gene interactions, as genes are constrained to regulate other genes via one or more TFs (figure 2D). For the inferred network 158 gene-gene interactions can be deduced from the 22 gene-TF and 84 TF-gene edges. However, literature mining reports only 5 gene-gene relationships between the 83 genes in the network, which is not nearly enough for validation purposes. Since the biological role of many selected genes remains to be investigated, we assume that the lack of text mining information is mainly due to the literature bias, by which genes that have been intensively studied for many years (e.g. TNF-alpha) are cited more often than less prominent genes.

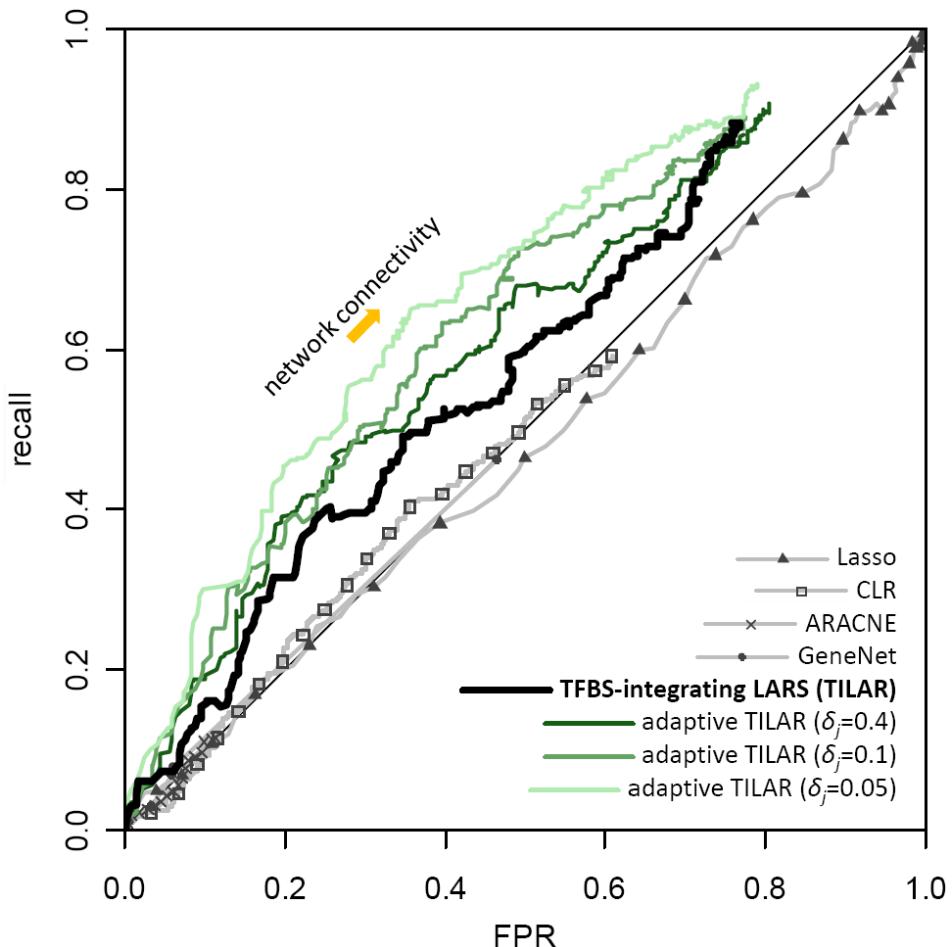
To overcome this issue, we sought for genes well described in the literature. For them we could expect many known gene regulatory interactions, so that a systematic evaluation of the performance of network reconstructions becomes feasible. We finally chose genes that are most frequently co-mentioned in the context of RA in PubMed. A respective list of genes was obtained from the Autoimmune Disease Database (version 1.2 as of August 19, 2008), which is a literature-based database that provides gene-disease associations of all known or suspected autoimmune diseases [45]. Out of the top 50 genes cataloged for the disease term "rheumatoid arthritis", 42 genes were measured in the Affymetrix dataset (additional file 6). We will denote the network of these 42 genes as the benchmarking GRN in the following.

Genes in the benchmarking network include several matrix-metallo-proteinases and a vast number of cytokines, in particular interleukins and the therapeutic target TNF-alpha. Overall, 389 gene-gene interactions between these genes could be retrieved through text mining. These interactions constitute a text mining network in which all but two genes are connected. The genes with the most connections are IL-6 (37), TNF-alpha (37) and IL-1 (33). We analyzed the regulatory regions of all the 42 genes and found overrepresented DNA-binding sites of 10 TFs (additional file 7). Amongst others, TFBS of NF-kappaB and AP-1 are significantly enriched, which is not sur-

prising as both TF complexes play central roles in immune regulation and are proven to be involved in the pathogenesis of RA [10,46]. In the resulting GRN template these 10 TFs are linked to 31 genes by means of 67 TF-gene interactions. When constructing a linear model of the benchmarking GRN using our novel inference algorithm TILAR, 13 TF-gene interactions were discarded during the backward stepwise selection procedure, i.e. 54 TF-gene interactions remained in the model. LARS then provided model predictions for different degrees of network connectivity (in successive LARS steps representing the dependency on parameter  $s$ ).

Next, we tested whether the inferred edges between genes exist or not in the text mining network containing 389 gene-gene interactions. For this purpose, we calculated the measures recall, precision and false positive rate (FPR) for different network connectivities. A plot of the precision versus the recall performance of a method (in case of LARS as a function of  $s$ ) and the ROC (receiver operating characteristic) curve, where recall is plotted against FPR, are two widely used visualizations for performance evaluation [43,44]. The ROC analysis allows comparison of the inference quality against a random prediction by calculating the area under the curve (AUC), while an AUC(ROC) close to 0.5 corresponds to a random forecast.

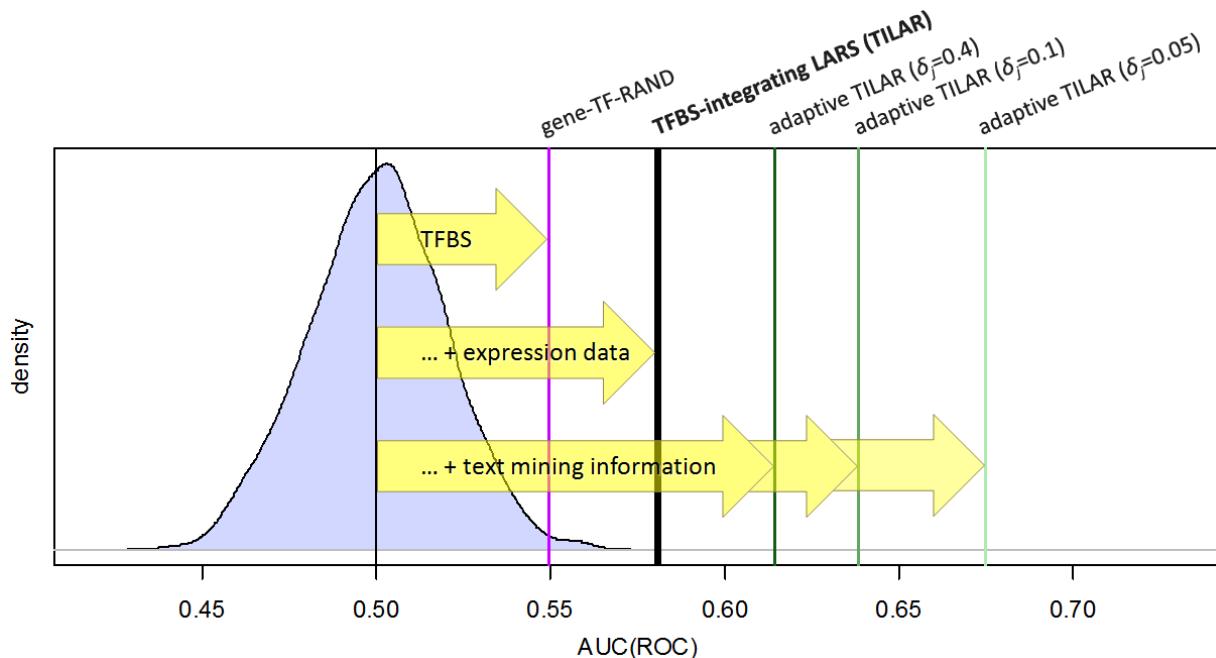
We utilized both recall-precision and ROC curves to assess and compare the performance of our algorithm and four different popular GRN inference methods: the conventional Lasso approach, CLR [47], ARACNE [48], and GeneNet [49] (see methods). While CLR and ARACNE use mutual information, GeneNet computes a partial correlation network. The resulting performance curves show that the proposed TILAR algorithm outperforms the other modeling algorithms (figure 6, additional file 8). When using AUC(ROC) as a single metric for benchmarking, the applied methods score as follows: Lasso – 0.478, ARACNE – 0.500, GeneNet – 0.503 and CLR – 0.504, whereas TILAR achieves an AUC(ROC) of 0.581. Next, we checked whether the algorithms performed significantly better than a random GRN prediction (RAND, see methods). We found, that the predictions of our approach were significantly better than RAND at the level  $\alpha = 0.05$  ( $P$ -value = 1.674e-05), while this was not the case for CLR, ARACNE, GeneNet and Lasso. Interestingly, gene-TF-RAND, another random algorithm that predicts gene regulatory interactions by including all 67 putative TF-gene interactions (i.e. the prior knowledge) into the model without considering the gene expression data (see methods), also yields a relatively high AUC(ROC) of 0.549 ( $P$ -value = 0.006). This suggests that TILAR performs well because of both the quality of TFBS predictions and data-fitting using LARS (figure 7).

**Figure 6**

**ROC curves for the benchmarking gene regulatory network.** The better a method performs, the closer its curve will be to the upper-left corner. The black curve represents the rating of our method when including 54 of 67 predicted TF-gene interactions. Remarkably, TILAR not only outperforms CLR, ARACNE, GeneNet and the conventional Lasso, but can also be combined with the adaptive LARS if adequate prior knowledge on gene-TF interactions is available. Using both techniques in combination we could infer gene-gene relationships more reliably.

Nevertheless, the AUC(ROC) of the TILAR method is still rather low. In our opinion this is not a general weakness of the modeling, but due to the fact that the information we used for model validation was obtained by text mining. This information is therefore incomplete and error-prone. A drawback is that the text mining network was constructed by searching through all biological literature and not only RA specific literature. Besides, text mining is obviously inappropriate to assess so far unknown regulatory interactions. In fact, the GRN model now provides new hypotheses that may be tested experimentally. However, there are several other factors that impede an accurate GRN reconstruction or an adequate performance

evaluation. First, regulatory networks can exhibit large dynamic topological changes [50]. Thus, among the interactions in our network we might only identify the most robust ones or those that are most relevant in the specific study, implying that some other could be missed, even if they have been biologically demonstrated. Second, the contribution of different cell types is lost in the study. Third, the text mining network contains undirected gene-gene interactions. In contrast, the proposed modeling approach assigns directed interactions between genes and TFs, i.e. gene-gene interactions are only implicitly defined in the model. Fourth, the network model might be too simple to reliably infer more complex interactions. Here,

**Figure 7**

**Performance gain using integrative modeling.** The expected AUC(ROC) of a random prediction is normally distributed around 0.5 as calculated by 1,000 repeated runs of RAND. The gene-TF-RAND algorithm considers the GRN template information, but assigns gene-TF interactions randomly. In contrast, the TILAR algorithm utilizes gene expression data to infer gene-TF interactions, while only a subset of the predicted TF-gene interactions is included into the final model. This significantly increases the inference quality. However, the method could be further improved by considering text mining information on which genes possibly regulate TF activity (adaptive TILAR). The combined inference method allows to strike a balance between data-fit and confidence in such putative gene-TF interactions by means of the parameters in  $\delta$ .

we assumed that gene regulatory effects are linear and additive, and precluded auto-regulation, i.e. a gene is not allowed to control a TF for which it possesses a TFBS. The latter because the data were not well suited for inferring a dynamic model. Moreover, the inference algorithm is based on co-expression at the transcriptional level, even if the amount of mRNA may not correspond to the level and (regulatory) activity of the proteins. However, more data would be required to infer more accurate models.

#### Adaptive TILAR – combined use of two techniques

Until now, we have shown that the proposed modeling approach, which utilizes gene expression data as well as TFBS predictions, performs fairly well in the reconstruction of GRNs. However, the method is not only an alternative to the adaptive LARS, but can also be used in combination with it (adaptive TILAR). The adaptive LARS [30] specified in equation (5) penalizes the coefficients  $\beta$  of equation (8) with weights  $\delta$ , dependent on whether the coefficient receives prior knowledge. In this way, we are able to integrate prior knowledge on gene-TF interactions as well. The lower we set the weights  $\delta$  for the coefficients

$\beta_j$  that represent putative gene-TF interactions, the more these interactions are *a priori* preferred to be in the model. The weights  $\delta$  thus allow to trade off data-fit and confidence in prior knowledge. However, accurate knowledge on gene-TF edges is difficult to obtain due to their variable meanings. For instance, intermediary molecules may account for such relationships. Here, we again applied text mining to retrieve potential gene-TF links (see methods). This way, we found 71 gene-TF relationships for the benchmarking GRN (e.g. the well-known activation of NF-kappaB by IL-1). We then evaluated the adaptive TILAR algorithm with three different weights for the preferred coefficients. As a result, the inference quality increased considerably (figure 6 and 7). When setting  $\delta_j = 0.4$  for the 71 preferred coefficients we obtained an AUC(ROC) of 0.615 ( $P\text{-value} = 1.768\text{e-}09$ ), for  $\delta_j = 0.1$  a value of 0.639 ( $P\text{-value} = 4.839\text{e-}13$ ), and for a very low  $\delta_j = 0.05$  a value of 0.675 ( $P\text{-value} < 2.2\text{e-}16$ ). An even lower  $\delta_j$  did not improve the result much. Thus, we can use information on TF-gene and gene-TF interactions to infer a GRN model, that predicts regulatory interactions between genes more reliably. Nevertheless, the true use of the com-

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

bination of adaptive and TFBS-integrating LARS requires more investigation. For instance, the determination of the weights  $\delta$  is not straightforward, as we might set different weights for each gene-TF interaction and even relatively high weights (i.e.  $\delta > 1.0$ ) if certain gene-TF relationships can be excluded *a priori*. However, this is beyond the scope of this paper.

### Conclusion

We developed TILAR – a method for deriving transcriptional regulatory networks from gene expression data by integrating TF binding predictions. The algorithm is also able to incorporate prior knowledge on the putative regulators of TF activity (adaptive TILAR). Our linear, additive modeling approach distinguishes genes and TFs in the network, and identifies the connections between them based on the fast LARS regression algorithm and specific constraints on the network structure. The major advantage of this modeling strategy is that only few model parameters are sufficient for a complex network, which is still easy to interpret. When applied on short-term gene expression profiles of RA patients treated with etanercept, the method uncovers molecular immunotherapeutic effects and thus provides testable hypotheses about the drugs' mechanisms of action. A closer look on the model revealed genes co-regulated by a common TF and TFs that regulate functionally related genes. Moreover, the reconstructed GRN exhibits a scale-free, self-regulating and massively parallel architecture.

We evaluated the inference quality using a text mining network and found that our modeling method outperforms all other algorithms tested. Notably, TILAR allows for a higher prediction accuracy than using just gene expression data or TF binding information alone. More efforts are needed to study different configurations of TILAR, e.g. we could analyze a larger DNA region for overrepresented TFBS, and to assess the benefit of combining this method with the adaptive LARS. Besides, further experiments need to be performed to verify specific interactions that were predicted by the model. However, even if significant theoretical and experimental challenges remain, we could demonstrate that organizing heterogeneous data and prior biological knowledge in systems biological models can strongly support the investigation of autoimmune diseases and their therapies. Supplementary materials including R codes are available at <http://www.hki-jena.de/index.php/0/2/490>.

### Methods

#### DNA microarray data pre-processing

We used the human DNA microarray dataset from Koczan *et al.* [13] including expression profiles of 19 etanercept-treated RA patients. Blood samples were taken for each patient before treatment as well as 72 and 144 hours after

first application of etanercept. Transcriptional levels of PBMC were then measured using Affymetrix Human Genome U133A arrays. As for 2 patients the third time-point is missing, the dataset consists of 55 microarray experiments. In the applied Affymetrix microarrays most probesets include probes matching transcripts from more than one gene and probes which do not match any transcribed sequence. Therefore, we utilized a custom chip definition file (CDF), that is based on the information contained in the GeneAnnot database [15,51]. GeneAnnot-based CDFs are composed of probesets including only probes matching a single gene and thus allow for a more reliable determination of expression levels. We used version 1.4.0 of the custom CDF and the MAS5.0 algorithm to pre-process the raw probe intensities. Data normalization was performed by a loess fit to the whole data with  $span = 0.05$  (using R package affy). Finally, the data processing yields mRNA abundances of 11,174 different genes.

#### Filtering differentially expressed genes

The filtering aims to identify a subset of genes significantly up- or down-regulated within the first week of therapy. A widely used filter criterion is the (logarithmized) fold-change from baseline. However, a fixed fold-change threshold ignores the inherent structure of DNA microarray data. Therefore, we applied an MA-plot-based signal intensity-dependent fold-change criterion (MAID filtering) to select genes. The MAID filtering takes into account that the variability in the log fold-changes increases as the measured signal intensity decreases [52]. First, the filtering procedure calculates for each gene the values A and M, which are commonly used for visualizing microarray data in an MA-plot. A is the log signal intensity of a gene averaged over all patients, while M is the mean intensity log-ratio between the baseline levels and the expression levels at day 3 and 6, respectively. Then, the intensity-dependent variability in the data is estimated by computing the interquartile range (IQR) of the M values in a sliding window. Afterwards, an exponential function  $f(x) = a \cdot e^{-bx} + c$  is fitted to the IQR's by a non-linear robust regression, which in turn is used to calculate so-called MAID-scores by dividing each M value by  $f(A)$ . As a consequence, the absolute value of a gene's MAID-score is higher, the more its expression level is altered after start of therapy. Furthermore, we assessed which genes are differentially expressed according to a paired *t*-test comparing the expression levels at day 3 and 6 versus baseline, respectively. Finally, we selected the genes having  $|MAID\text{-score}| > 2.5$  and *t*-test *P*-value  $< 0.05$ .

#### GO analysis

Overrepresented GO terms were found using GOSTATS, a Bioconductor package written in R. Each GO term is tested whether it is significantly associated to the list of filtered

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

genes out of the 11,174 measured genes. The analysis was performed for gene functional annotations of the biological process GO category.

#### **Identification of TF-gene interactions (GRN template)**

TFBS were derived from the UCSC database build hg18 [53]. The database provides a TFBS conserved (tfbsConsSites) track, that contains the location and score of TFBS conserved in the human/mouse/rat alignment. The data are purely computational and were generated using position weight matrices (PWMs) for TFBS contained in the public Transfac Matrix and Factor databases created by Biobase. For the whole human genome 3,837,187 TFBS predictions associated to 258 different PWMs (184 unique TF identifiers) can be found in the tfbsConsSites track. We defined the regulatory region of each gene as the 1,000 bp up- and downstream of the TSS (as stated in GeneCards database 2.38). This specification is in agreement with current findings by the ENCODE pilot project which revealed that regulatory sequences are symmetrically distributed around the TSS with no bias towards upstream regions [54]. Then, we scanned the regulatory regions of the selected genes for overrepresented TFBS. In doing so, each TF is tested whether its binding site occurs in this region for more genes than would be expected by chance. To take into account the inherent redundancy of the Transfac database, a TF is supposed to regulate a gene (TF-gene interaction) if any PWM for this TF matches the DNA sequence at the gene's regulatory region. Using a hypergeometric test analyzing the TF binding predictions for all the 11,174 genes measured, we can identify a subset of TFs associated to the genes in the network at the significance level  $\alpha = 0.05$ . This leads to a list of predicted TF-gene interactions that can serve as a template for GRN modeling.

#### **TFBS-integrating GRN inference (TILAR algorithm)**

First, the expression levels of each gene were standardized so that they have variance 1 and mean 0. Given these data, we then defined a regression equation according to equation (8), while considering the full set of putative TF-gene interactions. Afterwards, we calculated all LARS estimates (steps) for this equation using the R package lars with default settings. Each LARS estimate specifies a subset of covariates, i.e. states which gene-TF interactions are present in the model and which are not (in the latter case the corresponding model parameter is set to zero). To select a single estimate, we chose the model that minimizes Mallows' Cp statistic [55], thereby preventing overfitting and ensuring sparseness. The whole procedure was then repeated in a backward stepwise selection scheme in which TF-gene interactions were iteratively eliminated (or reinserted) if this allowed for a model that exhibits a smaller residual sum of squares (RSS). In this way, a subset of TF-gene interactions was found. For the regression

equation including this subset all possible Lasso estimates (see equation (4)) are provided by LARS. We then calculated for each LARS step the OLS fit using only the respective covariates (LARS/OLS hybrid [27]). Hence, we used LARS for variable selection, but not to estimate the model coefficients. Moreover, we evaluated the 10-fold cross-validation error ( $CV_{error}$ ) for each LARS/OLS solution and finally selected the most parsimonious model within 1 standard deviation from the  $CV_{error}$  minimum (additional file 4). It should be noted that we used the Cp statistic as a crude selection criterion during the backward stepwise selection procedure, because the Cp is much faster to compute than the  $CV_{error}$ .

To integrate prior knowledge on gene-TF interactions (as we did for the benchmarking GRN) we strictly followed the above learning strategy, except that we employed the adaptive variant of the LARS algorithm according to Zou [30]. The adaptive LARS assigns weights to each coefficient as written in equation (5). We penalized coefficients  $\beta_j$  for which we have no prior knowledge with a neutral weight  $\delta_j = 1.0$ . If literature mining suggested a gene-TF interaction we penalized the corresponding coefficient with a smaller  $\delta_j$  (0.4, 0.1 and 0.05, respectively) to improve variable selection.

The learning strategy of the (adaptive) TILAR is summarized as follows:

1. Define  $D$  as the given (standardized) gene expression data
2. Define  $P$  as the given set of putative TF-gene interactions (GRN template)
3. Use  $D$  to specify regression equation  $L$  subject to  $P$  according to equation (8)
4. Solve  $L$  using (adaptive) LARS and calculate  $RSS(Cp_{min})$ , i.e. the RSS of the LARS estimate that minimizes Cp
5. Optional: Perform a backward stepwise selection on  $P$ , i.e. iteratively and exhaustively remove or reinsert elements in  $P$  and repeat 3. and 4., and stop when a local minimum for  $RSS(Cp_{min})$  is found
6. Recompute the regression coefficients to  $L$  in terms of a LARS/OLS hybrid and return the most parsimonious estimate within 1 standard deviation of the 10-fold  $CV_{error}$  minimum

#### **Performance evaluation**

We used gene-gene interaction information for benchmarking. The software PathwayArchitect 2.0.1 was

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

employed to automatically extract such gene-gene links from the literature. We retrieved only gene-gene interactions of context type "expression" and "regulation" as labeled by PathwayArchitect. To obtain putative gene-TF links (which were used for the adaptive TILAR) we also considered interactions of type "protein modification". The gene-gene information was applied to assess the inference quality of our and a total of four other easy-to-apply GRN inference methods, namely CLR, ARACNE, GeneNet, and the conventional Lasso, while we did not take into account the directions of the relationships. CLR, ARACNE and GeneNet are thought to build (undirected) gene association networks and have been implemented by use of the R packages minet and GeneNet. To compute the entire set of Lasso solutions to equation (3) we used the LARS modification (R package lars). All methods were run on standardized gene expression levels with default settings. Moreover, a random inference algorithm called RAND was implemented, which randomly assigns connections between genes until a fully connected network is formed. The RAND method was further adapted to infer networks of TF-gene and gene-TF interactions similar to the proposed modeling scheme (gene-TF-RAND). More specifically, gene-TF-RAND utilizes all the TF-gene interactions predicted by the TFBS overrepresentation analysis and randomly adds gene-TF edges to the network. As for TILAR, gene-TF interactions were not allowed when gene and TF were already connected by a TF-gene interaction, and gene-gene links result implicitly. The AUC(ROC) value of gene-TF-RAND was obtained by the mean of 1,000 repeated runs. Apart from that, we tested whether any inference technique performed significantly better than a random prediction. For this purpose, we calculated P-values which specify the probability that an AUC(ROC) value computed by RAND will be higher than the AUC(ROC) value of the particular inference algorithm. The P-values are calculated by 1 minus the cumulative probabilities, which are evaluated at the AUC(ROC) value of the respective method, of the normal distribution having the mean and standard deviation of 1,000 RAND-calculated AUC(ROC) values.

## Abbreviations

TF: transcription factor; GRN: gene regulatory network; RA: rheumatoid arthritis; TFBS: TF binding site; LARS: least angle regression; OLS: ordinary least squares; GO: gene ontology; TILAR: TFBS-integrating LARS; PBMC: peripheral blood mononuclear cells; MAID: MA-plot-based signal intensity-dependent fold-change criterion; TSS: transcription start site; Lasso: least absolute shrinkage and selection operator; FPR: false positive rate; ROC: receiver operating characteristic; AUC: area under the curve; CDF: chip definition file; IQR: interquartile range; PWM: position weight matrix; RSS: residual sum of squares.

## Conflict of interests

The authors declare that they have no competing interests.

## Authors' contributions

RGu and H-JT directed the study. MH carried out the analyses on the data and wrote the paper. RGu, RGo and RE assisted in interpretation of the results and contributed to writing the paper. All authors read and approved the final manuscript.

## Additional material

### Additional file 1

List of 83 genes with significant expression changes during first week of therapy. The table provides diverse types of information for each gene, e.g. Entrez ID, official full name and the calculated MAID-scores.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-262-S1.xls>]

### Additional file 2

Filtering of genes regulated in response to etanercept therapy. (A) Superimposed MA-plot visualizing the applied gene filtering method. Here, gene expression levels measured 3 days after therapy onset are compared with baseline levels. The MAID filtering takes into account that the variability in the mean log-fold changes ( $M$ ) depends on the mean log signal intensity ( $A$ ). 37 genes showed an up- or down-regulation at day 3 (green). (B) In a similar manner, 57 genes were found higher or lower expressed at day 6 in comparison to baseline. In this way, 83 different genes were selected in total. (C) Mean time-courses of these 83 genes. 25 genes were found up- or down-regulated at day 3 (left), 45 at day 6 (middle) and 13 at day 3 and 6 (right).

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-262-S2.png>]

### Additional file 3

Overrepresented terms of the GO biological process ontology. P-values were computed for each GO term based on the hypergeometric distribution. Only functional categories with P-value < 0.01 and where at least 3 out of 83 genes are associated ("Count") are shown.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-262-S3.xls>]

### Additional file 4

Model selection using cross-validation. Training error (scaled by 10) and 10-fold  $CV_{error}$  (RSS mean of 10 subsets) are shown for the LARS/OLS solutions of the first 300 LARS steps. The blue area represents the standard deviation of  $CV_{error}$ . The red line shows the LARS step selected for the final model, i.e. the most parsimonious model within 1 standard deviation from the  $CV_{error}$  curve minimum, for which 22 model parameters are non-zero.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-262-S4.png>]

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

### Additional file 5

Zip-archived Cytoscape session file of the reconstructed GRN. The network model contains predicted regulatory interactions of genes responsive to etanercept therapy in RA. A simplified visualization of the network is shown in figure 3, while detail views are shown in figure 5.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-262-S5.zip>]

### Additional file 6

List of the 50 most frequently mentioned genes in the context of RA. 42 of these genes were measured in the dataset and used to evaluate the performance of our modeling approach.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-262-S6.xls>]

### Additional file 7

Overrepresented binding sites for the benchmarking gene regulatory network. 10 transcription factors represented by 20 Transfac binding profiles were found to be enriched, providing 67 predicted TF-gene interactions in total.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-262-S7.xls>]

### Additional file 8

Recall-precision curves for the benchmarking GRN. We evaluated the performance of different modeling strategies based on gene-gene relationships found by text mining. The black curve represents the rating of our method when including 54 of 67 predicted TF-gene interactions. The TILAR approach outperforms CLR, ARACNE, GeneNet and the conventional Lasso. When used in combination with the adaptive LARS we could further increase the inference quality.

[Click here for file](#)

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-262-S8.png>]

## Acknowledgements

We thank Peter Lorenz and Dirk Koczan for helpful discussions. Thanks to Steven John Smith for critical proofreading of the manuscript.

Funding: This study was supported by grants from the German Federal Ministry of Education and Research (BMBF, BioChancePlus, 0313692D).

## References

- Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** Nucleic Acids Res 2003, **31**(1):374-378.
- Filkov V: **Identifying Gene Regulatory Networks from Gene Expression Data.** In *Handbook of Computational Molecular Biology* Edited by: Aluru S. Chapman & Hall/CRC Press; 2005:27.1-27.29.
- Hecker M, Lambeck S, Toepfer S, van Someren EP, Guthke R: **Gene regulatory network inference: Data integration in dynamic models – A review.** Biosystems 2009, **96**(1):86-103.
- Cho KH, Choo SM, Jung SH, Kim JR, Choi HS, Kim J: **Reverse engineering of gene regulatory networks.** IET Syst Biol 2007, **1**(3):149-163.
- Werhli AV, Husmeier D: **Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge.** Stat Appl Genet Mol Biol 2007, **6**:Article 15.
- Benson M, Breitling R: **Network theory to understand microarray studies of complex diseases.** Curr Mol Med 2006, **6**(6):695-701.
- Kavanaugh AF, Lipsky P: **Rheumatoid arthritis.** In *Inflammation: Basic Principles and Clinical Correlates* 3rd edition. Edited by: Gallin JL, Snyderman R. Philadelphia, PA, USA: Lippincott Williams & Wilkins; 1999:1017-1037.
- McInnes IB, Schett G: **Cytokines in the pathogenesis of rheumatoid arthritis.** Nat Rev Immunol 2007, **7**(6):429-442.
- Brennan FM, Maini RN, Feldmann M: **Role of pro-inflammatory cytokines in rheumatoid arthritis.** Springer Semin Immunopathol 1998, **20**(1-2):133-147.
- Smolen JS, Steiner G: **Therapeutic strategies for rheumatoid arthritis.** Nat Rev Drug Discov 2003, **2**(6):473-488.
- Feldmann M, Maini RN: **Lasker Clinical Medical Research Award. TNF defined as a therapeutic target for rheumatoid arthritis and other autoimmune diseases.** Nat Med 2003, **9**(10):1245-1250.
- Glocker MO, Guthke R, Kekow J, Thiesen HJ: **Rheumatoid arthritis, a complex multifactorial disease: on the way toward individualized medicine.** Med Res Rev 2006, **26**(1):63-87.
- Koczan D, Drynda S, Hecker M, Drynda A, Guthke R, Kekow J, Thiesen HJ: **Molecular discrimination of responders and non-responders to anti-TNFalpha therapy in rheumatoid arthritis by etanercept.** Arthritis Res Ther 2008, **10**(3):R50.
- Fernald GH, Knott S, Pachner A, Caillier SJ, Narayan K, Oksenberg JR, Mousavi P, Baranzini SE: **Genome-wide network analysis reveals the global properties of IFN-beta immediate transcriptional effects in humans.** J Immunol 2007, **178**(8):5076-5085.
- Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli GA, Bicciato S: **Novel definition files for human GeneChips based on GeneAnnot.** BMC Bioinformatics 2007, **8**:446.
- Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** Nucleic Acids Res 2005, **33**(20):e175.
- Boissier MC, Assier E, Biton J, Denys A, Falgarone G, Bessis N: **Regulatory T cells (Treg) in rheumatoid arthritis.** Joint Bone Spine 2009, **76**(1):10-14.
- Peng G, Guo Z, Kiniwa Y, Voo KS, Peng W, Fu T, Wang DY, Li Y, Wang HY, Wang RF: **Toll-like receptor 8-mediated reversal of CD4+ regulatory T cell function.** Science 2005, **309**(5739):1380-1384.
- Akira S, Isshiki H, Sugita T, Tanabe O, Kinoshita S, Nishio Y, Nakajima T, Hirano T, Kishimoto T: **A nuclear factor for IL-6 expression (NF-IL6) is a member of a C/EBP family.** EMBO J 1990, **9**(6):1897-1906.
- Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E: **Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters.** Gene 2007, **389**(1):52-65.
- D'haeseleer P, Wen X, Fuhrman S, Somogyi R: **Linear modeling of mRNA expression levels during CNS development and injury.** Pac Symp Biocomput 1999, **4**:41-52.
- van Someren EP, Vaes BL, Steegenga WT, Sijbers AM, Decherig KJ, Reinders MJ: **Least absolute regression network analysis of the murine osteoblast differentiation network.** Bioinformatics 2006, **22**(4):477-484.
- Yong-A-Poi J: **Adaptive least absolute regression network analysis improves genetic network reconstruction by employing prior knowledge.** In Master's thesis Delft University of Technology, Department of Mediamatics; 2008.
- Thorsson V, Hörnquist M, Siegel AF, Hood L: **Reverse engineering galactose regulation in yeast through model selection.** Stat Appl Genet Mol Biol 2005, **4**:Article 28.
- Guthke R, Möller U, Hoffmann M, Thies F, Töpfer S: **Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection.** Bioinformatics 2005, **21**(8):1626-1634.
- Tibshirani R: **Regression shrinkage and selection via the lasso.** J Royal Statist Soc B 1996, **58**:267-288.

BMC Bioinformatics 2009, 10:262

<http://www.biomedcentral.com/1471-2105/10/262>

27. Efron B, Hastie T, Johnstone I, Tibshirani R: **Least angle regression.** *Ann Statist* 2004, **32**(2):407-499.
28. van Someren EP, Wessels LFA, Reinders MJT, Backer E: **Regularization and noise injection for improving genetic network models.** In *Computational And Statistical Approaches To Genomics* Edited by: Zhang W, Shmulevich I. Boston, MA, USA: Kluwer Academic Publishers; 2002:211-226.
29. Gustafsson M, Hörnquist M, Lombardi A: **Constructing and analyzing a large-scale gene-to-gene regulatory network – lasso-constrained inference and biological validation.** *IEEE/ACM Trans Comput Biol Bioinform* 2005, **2**(3):254-261.
30. Zou H: **The adaptive lasso and its oracle properties.** *J Am Stat Assoc* 2006, **101**:1418-1429.
31. Kitano H: *Foundations of Systems Biology* Cambridge, MA, USA: MIT Press; 2001.
32. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells.** *Nat Genet* 2005, **37**(4):382-390.
33. Jordan IK, Mariño-Ramírez L, Wolf YI, Koonin EV: **Conservation and coevolution in the scale-free human gene coexpression network.** *Mol Biol Evol* 2004, **21**(11):2058-2070.
34. Clauset A, Shalizi CR, Newman MEJ: **Power-law distributions in empirical data.** 2007 [<http://arxiv.org/abs/0706.1062>].
35. Ishibashi K, Kuwahara M, Gu Y, Tanaka Y, Marumo F, Sasaki S: **Cloning and functional expression of a new aquaporin (AQP9) abundantly expressed in the peripheral leukocytes permeable to water and urea, but not to glycerol.** *Biochem Biophys Res Commun* 1998, **244**(1):268-274.
36. Strubin M, Newell JW, Matthias P: **OBF-1, a novel B cell-specific coactivator that stimulates immunoglobulin promoter activity through association with octamer-binding proteins.** *Cell* 1995, **80**(3):497-506.
37. Rey M, Vicente-Manzanares M, Viedma F, Yáñez-Mó M, Urzainqui A, Barreiro O, Vázquez J, Sánchez-Madrid F: **Cutting edge: association of the motor protein nonmuscle myosin heavy chain-IIA with the C terminus of the chemokine receptor CXCR4 in T lymphocytes.** *J Immunol* 2002, **169**(10):5410-5414.
38. O'Hara R, Murphy EP, Whitehead AS, Fitzgerald O, Bresnihan B: **Acute-phase serum amyloid A production by rheumatoid arthritis synovial tissue.** *Arthritis Res* 2000, **2**(2):142-144.
39. Ogura Y, Inohara N, Benito A, Chen FF, Yamaoka S, Nunez G: **Novel, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-κappaB.** *J Biol Chem* 2000, **276**(7):4812-4818.
40. Salmi M, Koskinen K, Henttinen T, Elimä K, Jalkanen S: **CLEVER-I mediates lymphocyte transmigration through vascular and lymphatic endothelium.** *Blood* 2004, **104**(13):3849-3857.
41. Qian F, Kruse U, Lichter P, Sippel AE: **Chromosomal localization of the four genes (NFIA, B, C, and X) for the human transcription factor nuclear factor I by FISH.** *Genomics* 1995, **28**(1):66-73.
42. Goertsches RH, Hecker M, Zettl UK: **Monitoring of multiple sclerosis immunotherapy: From single candidates to biomarker networks.** *J Neurol* 2008, **255**(S6):48-57.
43. Stolovitzky G, Monroe D, Califano A: **Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference.** *Ann N Y Acad Sci* 2007, **1115**:1-22.
44. Stolovitzky G, Prill RJ, Califano A: **Lessons from the DREAM2 Challenges.** *Ann N Y Acad Sci* 2009, **1158**:159-195.
45. Karopka T, Fluck J, Mevissen HT, Glass A: **The Autoimmune Disease Database: a dynamically compiled literature-derived database.** *BMC Bioinformatics* 2006, **7**:325.
46. Kinne RW, Bräuer R, Stuhlmüller B, Palombo-Kinne E, Burmester GR: **Macrophages in rheumatoid arthritis.** *Arthritis Res* 2000, **2**(3):189-202.
47. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.** *PLoS Biol* 2007, **5**(1):e8.
48. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** *BMC Bioinformatics* 2006, **7**(Suppl 1):S7.
49. Opgen-Rhein R, Strimmer K: **From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data.** *BMC Syst Biol* 2007, **1**:37.
50. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes.** *Nature* 2004, **431**(7006):308-312.
51. Chalifa-Caspi V, Shmueli O, Benjamin-Rodrig H, Rosen N, Shmoish M, Yanai I, Ophir R, Kats P, Safran M, Lancet D: **GeneAnnot: interfacing GeneCards with high-throughput gene expression compendia.** *Brief Bioinform* 2003, **4**(4):349-360.
52. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3**(11):research0062.
53. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2009.** *Nucleic Acids Res* 2009:D755-761.
54. ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**(7146):799-816.
55. Mallows C: **Some comments on cp.** *Technometrics* 1973, **15**:661-675.

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)



## 5. Manuscript III

---

### **Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN-beta-1b treatment in relapsing remitting MS**

Robert Hermann Goertsches, Michael Hecker, Dirk Koczan,  
Pablo Serrano-Fernández, Steffen Möller, Hans-Jürgen Thiesen,  
and Uwe Klaus Zettl

*Pharmacogenomics* 2010, 11(2):147-161.

---



## RESEARCH ARTICLE

For reprint orders, please contact: [reprints@futuremedicine.co](mailto:reprints@futuremedicine.co)

## Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN- $\beta$ -1b treatment in relapsing remitting MS

**Aims:** In multiple sclerosis patients, treatment with recombinant IFN- $\beta$  (rIFN- $\beta$ ) is partially efficient in reducing clinical exacerbations. However, its molecular mechanism of action is still under scrutiny. **Materials & methods:** We used DNA microarrays (Affymetrix, CA, USA) and peripheral mononuclear blood cells from 25 relapsing remitting multiple sclerosis patients to analyze the longitudinal transcriptional profile within 2 years of rIFN- $\beta$  administration. Sets of differentially expressed genes were attained by applying a combination of independent criteria, thereby providing efficient data curation and gene filtering that accounted for technical and biological noise. Gene ontology term-association analysis and scientific literature text mining were used to explore evidence of gene interaction. **Results:** Post-therapy initiation, we identified 42 (day 2), 175 (month 1), 103 (month 12) and 108 (month 24) differentially expressed genes. Increased expression of established IFN- $\beta$  marker genes, as well as differential expression of circulating IFN- $\beta$ -responsive candidate genes, were observed. *MS4A1* (CD20), a known target of B-cell depletion therapy, was significantly downregulated after one month. *CMPK2*, *FCER1A*, and *FFAR2* appeared as hitherto unrecognized multiple sclerosis treatment-related differentially expressed genes that were consistently modulated over time. Overall, 84 interactions between 54 genes were attained, of which two major gene networks were identified at an earlier stage of therapy: the first ( $n = 15$  genes) consisted of mostly known IFN- $\beta$ -activated genes, whereas the second ( $n = 12$ ) mainly contained downregulated genes that to date have not been associated with IFN- $\beta$  effects in multiple sclerosis array research. **Conclusion:** We achieved both a broadening of the knowledge of IFN- $\beta$  mechanism-of-action-related constituents and the identification of time-dependent interactions between IFN- $\beta$  regulated genes.

**KEYWORDS:** autoimmunity ■ DNA microarray ■ IFN- $\beta$  ■ immunomodulatory drug  
■ multiple sclerosis ■ network analysis ■ peripheral blood

Multiple sclerosis (MS) is a chronic, inflammatory, disabling disease of the CNS, and is the most frequent disorder that causes persistent deficits [1]. Since the early 1990s, disease-modifying drugs have played a major role in MS treatment, of which the most applied are IFN- $\beta$ , glatiramer acetate and monoclonal antibodies [2]. Currently, the interferons are the most distinguished when undertaking an immunomodulatory therapy [3]; however, they are only partially effective, and precise molecular mechanisms remain unclear. Deciphering the pharmacogenomic effects of recombinant IFN- $\beta$  (rIFN- $\beta$ ) treatment has proven to be a challenge, but integrated efforts (e.g., between molecular biology, neurology and bioinformatics) shall eventually lead to the elucidation of the underlying complex processes that involve multiple genes [4,5].

The use of divergent experimental approaches based on transcriptomics and high-content microarrays has resulted in the identification of various biological IFN- $\beta$  response markers: for example, *MX1*, *OAS* and other interferon-stimulated genes

(ISGs) [6–9]. Hence, genome-wide hypothesis-free expression analysis of IFN- $\beta$  treated peripheral blood cells can document the broad effects of the drug well [10–26]. Huge amounts of quantitative data and lists of differentially expressed genes (DEGs) were thereby generated, with anticipated IFN- $\beta$ -induced genes being redundantly reported [27].

It is generally agreed that rIFN- $\beta$  administration ameliorates immune dysfunction, which is a dynamic multicomponent process covering the fields of immune cell regulation such as inhibition of T-cell proliferation, reduction of pathologic blood–brain endothelium permeability, and respective T-cell transmigration via interference with cell adhesion and upregulation of anti-inflammatory cytokines. The main cascades induced represent antiviral activity, chemotaxis, apoptosis, antigen presentation, Th1 differentiation, humoral immunity and dendritic cell maturation [28]. In the search for biomarkers, it became apparent that for several genes the sensitivity to rIFN- $\beta$  appeared to be

Robert H Goertsches<sup>1,2†</sup>,  
Michael Hecker<sup>2</sup>, Dirk  
Koczan<sup>3</sup>, Pablo Serrano-  
Fernandez<sup>3</sup>, Steffen  
Moeller<sup>3</sup>, Hans-Juergen  
Thiesen<sup>3</sup> & Uwe K Zettl<sup>1</sup>

<sup>1</sup>Author for correspondence:

<sup>2</sup>Department of Neurology,  
University of Rostock,  
Gehlsheimer Str. 20,  
18047 Rostock, Germany  
Tel.: +49 381 494 5891  
Fax: +49 381 494 5882  
[robert.goertsches@med.uni-rostock.de](mailto:robert.goertsches@med.uni-rostock.de)

<sup>3</sup>Leibniz Institute for Natural  
Product Research & Infection  
Biology – Hans Knöll Institute,  
Jena, Germany

<sup>3</sup>Institute of Immunology,  
University of Rostock,  
Schillingallee 70,  
18055 Rostock, Germany

future  
medicine part of fsg

## RESEARCH ARTICLE

Goertsches, Hecker, Koczan *et al.*

fairly variable between patients with MS [29]. To more comprehensively investigate the effects on transcriptional regulation, gene network inference techniques can be applied. Initial efforts were made to explore gene regulation in response to IFN- $\beta$  by incorporating network analysis, thereby effectively combining data- and knowledge-driven analysis [20,23,30].

In most transcriptomic studies that covered the earlier time window instantly after drug administration, a significant fraction of genes showed rapid increase from baseline. However, most of these changes reverted, as shown by prolonged transcriptomic analyses, over days [13,23–25], months [18] or years [17,26]. Therefore, as a complementary measure, for 2 years, we investigated the transcriptional effects of rIFN- $\beta$  1b treatment on peripheral mononuclear blood cells (PBMCs) at the end of the 48-h time window of the application regimen. The objective of this study was to examine the pharmacodynamic reaction of treated MS patients on the transcriptomic level, with focus on interaction structures between filtered genes (using PathwayArchitect™ information), as well as to highlight their biological functions (gene ontology [GO] terms) in time. The blood samples were obtained from all patients and were analyzed equally, following the principle of a nonhypothesis-/explorative-driven expression measurement. We furthermore deemed it important that the effects of one exclusive rIFN- $\beta$  therapy were investigated, as dose and route of administration act with distinctive effects [31]; if not on the substantially induced ISGs such as *MX1*, then more likely on the weaker induced ones. The outcome should provide opportunities to further understand IFN- $\beta$ 's mechanism of action at a molecular level, independently from any clinical response status, and to contemplate what effects are to be expected when combination regimens are at issue.

### Materials & methods

The study was approved by the University of Rostock's (Rostock, Germany) ethics committee and was carried out according to the Declaration of Helsinki. Informed consent from study participants was collected prior to study onset.

A total of 25 Caucasian patients (TABLE 1) with diagnosed relapsing remitting MS (RRMS) according to the McDonald criteria [32] were prescribed a first immunotherapy of rIFN- $\beta$ -1b. 250  $\mu$ g (8 MIU) of the drug were administered subcutaneously every other day. None of the patients had previously been medicated with

immunomodulatory or immunosuppressive agents or had ever received cytotoxic treatments, and all were free of glucocorticoid treatment for at least 30 days prior to blood extraction. 15 ml of peripheral venous EDTA treated blood were withdrawn prior to first and consecutive drug administrations, providing *ex vivo* material before (baseline) as well as 2 days (D2), 1 month (M1), 1 year (M12) and 2 years (M24) post-therapy initiation (PTI). Samples were always collected at the same time of day and processed within 1 h; the range of intervals of blood collection did not exceed 1 h.

### ■ RNA extraction & hybridization of Affymetrix HGU133 A & B microarrays

Total RNA of Ficoll-treated PBMCs from each sample were isolated following the manufacturer's protocol (RNeasy, Qiagen, Hilden, Germany). Initial RNA and final cRNA concentrations were determined spectrophotometrically by a Nanodrop® 1000 (Thermo Fisher Scientific, MA, USA), and quality control was performed by native ethidium bromide agarose gel electrophoresis. Samples of RNA (total 7  $\mu$ g) were labeled and hybridized according to the supplier's instructions (Affymetrix, CA, USA). The arrays, interrogating 44,928 probesets, were scanned at 3- $\mu$ m resolution using the GeneArray® Chip Scanner 2500 (Hewlett Packard, CA, USA).

### Validation of microarray data by means of real-time PCR

Quantitative real-time reverse transcription PCR (real-time PCR) was applied to confirm observed changes in gene expression. Transcript levels of 15 selected genes were measured in a subset of the samples. Details on the experimental procedure and on the analysis of the real-time PCR data are described in the ONLINE SUPPLEMENTARY MATERIAL 1 ([www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)).

### ■ Data analysis: data preprocessing, curation & filtering

Primary data analysis and quality control was carried out using the GeneChip® operating software (GCOS 1.4, Affymetrix) and the MAS5.0 (Microarray Suite 5.0, Affymetrix) statistical algorithms for probe level analysis. To discover transcriptional effects with regard to rIFN- $\beta$ -1b subcutaneous administration, we initially applied two criteria based on MAS5.0 labels to erase noninformative probesets (data curation) and, subsequently, three criteria that delivered an output of DEGs (gene filtering).

Cleaning the dataset of noninformative probesets was performed by removing probesets consistently labeled 'absent' in all 25 individuals, and removing probesets consistently labeled 'no change' in all 25 individuals (TABLE 1). One element of the subsequent filtering is described as follows. An accepted filtering criterion is the (log) fold-change from baseline; however, a fixed threshold ignores the inherent structure of DNA microarray data [33]. Therefore, we applied a MA-plot-based signal intensity-dependent fold-change criterion (MAID filtering) to identify strongly regulated probesets [34]. In brief, MAID considers that the variability in log fold-change increases as the measured signal intensity decreases [33]. Values 'A' and 'M' were calculated; 'A' represents the average log intensity for a probeset, while 'M' is the mean intensity log-ratio between baseline and the respective subsequent time point (FIGURE 1). The intensity-dependent variability was estimated by computing the interquartile range (IQR) for the M values in a moving window. An exponential function  $f(x) = a \times e^{bx} + c$  was then fitted to the IQRs by a nonlinear robust regression. By dividing each M value by  $f(A)$ , we calculated so-called MAID-scores. Finally, to filter significantly up- and downregulated probesets relative to baseline, we combined outlined MAID-scores, outcomes of a paired two-sample t-test, and increase and decrease of MAS5.0 generated labels [35] (FIGURE 2):  $|MAID\text{-score}| > 2$ ; statistical significance below  $p = 0.05$ ; minimum 50% of patients per probeset display increase and decrease, respectively.

#### Affymetrix probeset quality control applying GeneAnnot

Improvements in genome sequence annotation revealed discrepancies in the original probeset-gene assignment of Affymetrix microarrays. In the applied generation of Affymetrix human GeneChips, numerous probesets include probes matching transcripts from more than one gene and probes that do not match any transcribed sequence [36]. To remove such equivocal probesets from the data, we utilized a probeset specificity cutoff (0.7) that was based on the information contained in the GeneAnnot database version 1.7 [37].

#### Functional analysis (gene ontology term enrichment)

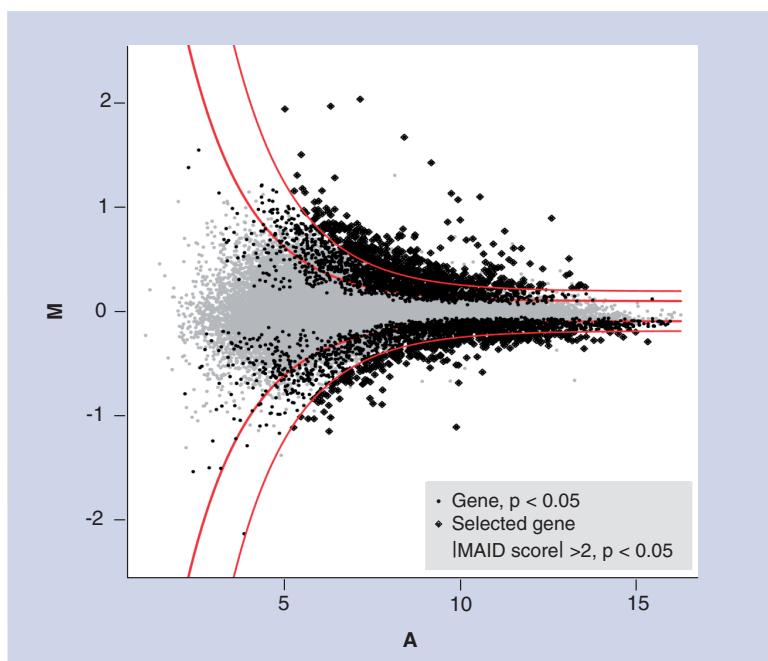
Functional annotation analyses based on the association of GO biological process terms [38] with filtered DEG sets were carried out by application

**Table 1. Demographic data of analyzed individuals.**

Variable	Value
Subjects (n)	25
Gender ratio (female:male)	16:9
Age at study, years (mean $\pm$ SD)	39.6 $\pm$ 10.9
Patients with EDSS Q25–Q75	2.2 (1.0–3.0)
Relapse rate (mean $\pm$ SD)	1.6 $\pm$ 0.9

EDSS: Expanded disability status scale; SD: Standard deviation.

of GOstats, which is a Bioconductor package written in R [39]. It allows the testing of GO terms for over-representation by computing a probability based on the hypergeometric distribution, which assesses whether the number of selected genes associated with the term is larger than that expected by chance. As a reference dataset, that is, the gene universe, we used all unique Entrez IDs ( $n = 12,377$ ) of HG-U133A/B.



**Figure 1. Superimposed MA plot visualizes the applied intensity dependent probeset filter on the microarray data.** MAID filtering of 44,928 probesets in response to IFN- $\beta$ -1b subcutaneous treatment considers the intensity-dependent variability in the log-fold changes; attained from peripheral mononuclear blood cell mRNA of relapsing remitting multiple sclerosis patients ( $n = 25$ ). Averaged expression intensities of probesets and respective changes from baseline to 48 h post-therapy initiation are depicted as grey points. 'A' represents the average log intensity for a probeset; 'M' the mean intensity log-ratio of the expression at the baseline and day 2. An exponential function is fitted to the shape of the data by a nonlinear robust regression (curved line, MAID regression curve). Hence, the MAID-score is low when the signal intensity of a probeset is marginal or not affected by recombinant IFN- $\beta$ -1b. The gene-filtering criteria,  $p < 0.05$  and  $|MAID\text{-score}| > 2$ , are introduced in this visualization (MAID score cutoff curve); probesets that withhold these are shown as diamonds. MAID: MA-plot-based signal intensity-dependent fold-change criterion.

## RESEARCH ARTICLE

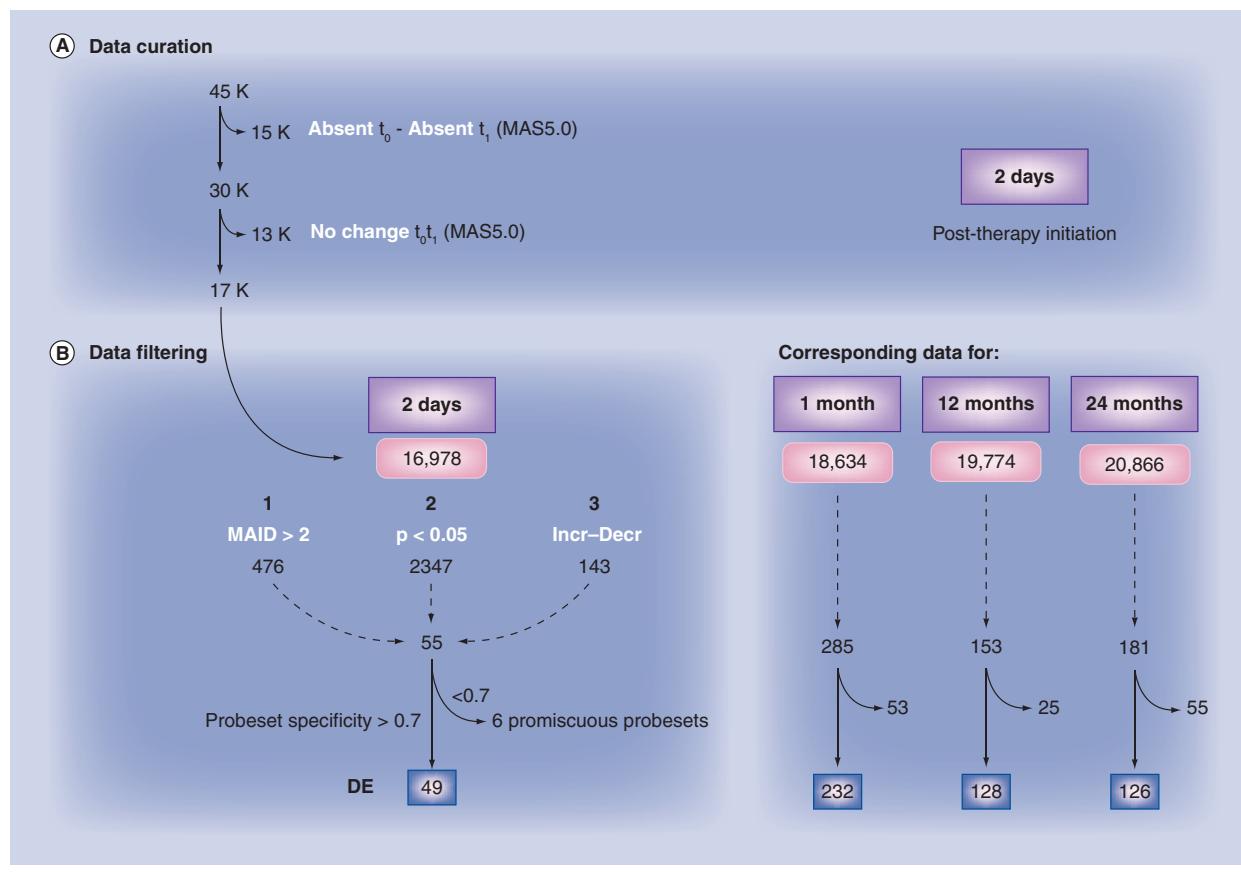
Goertsches, Hecker, Koczan *et al.*

Figure 2. Workflow and overview of data curation and filtering.

### ■ Knowledge-based network construction

In order to build networks of interacting genes, we placed our filtered DEGs in a systems perspective by consulting interaction information of the Pathway Architect (PA) 2.0.1 database (Stratagene Inc., CA, USA) [101]. We used biological relationships that PA distinguishes, such as ‘positive regulation = stimulation’ and ‘negative regulation = inhibition’, and basic connections ‘regulation’, ‘binding’, ‘expression’, ‘metabolism’, ‘transport’ and ‘protein modification’. The multiple interactions between genes were visualized using Cytoscape 2.6.1 [40].

### Results

Longitudinal RNA samples taken at baseline, 2 days, and 1, 12, and 24 months were collected from 25 patients with RRMS and were each hybridized onto Affymetrix HG-U133 microarray sets. The expression dataset of 250 A- and B-chip measurements have been deposited in the National Center for Biotechnology

Information Gene Expression Omnibus (GEO; [102]) and are accessible through GEO Series accession number (in process).

### ■ Filtering differentially expressed genes

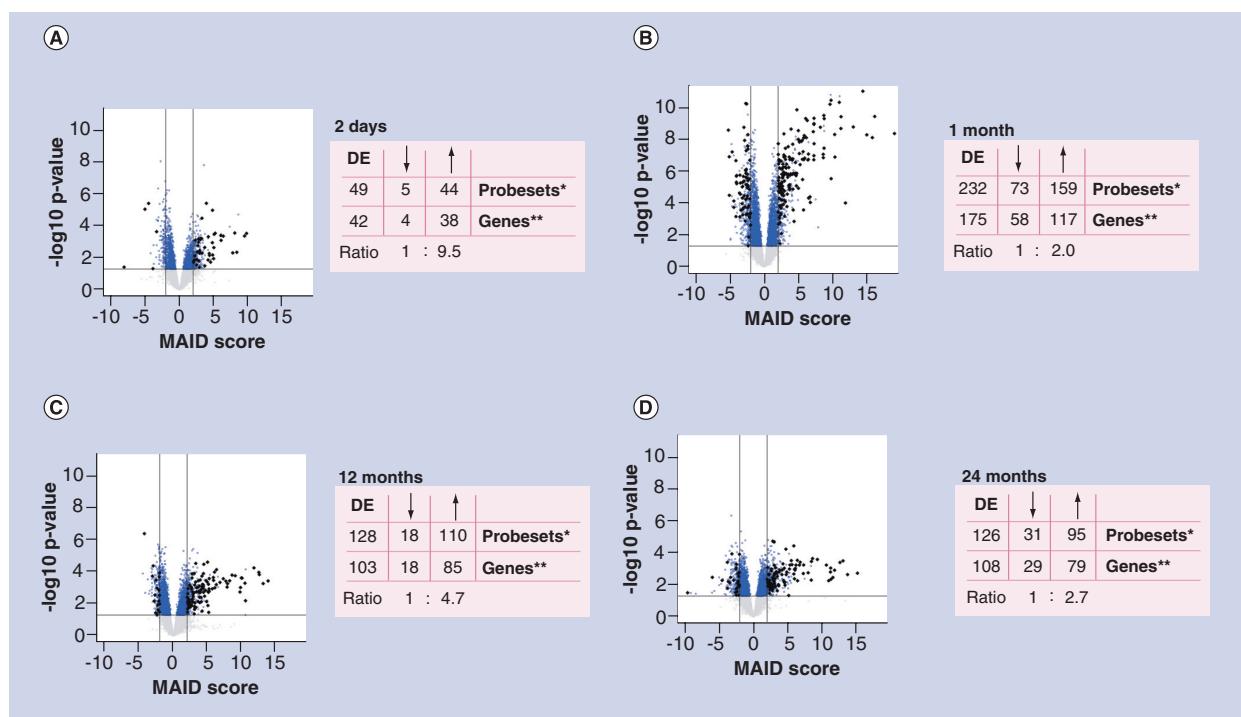
To select genes showing significant transcript changes in response to rIFN-β in PBMCs, we applied five filtering criteria (see ‘Methods’ section; FIGURES 1 & 2). Exemplarily, at D2 PTI, data curation sieved both nonexpressed and noninformative probesets, the latter representing genes that were expressed but not changed owing to rIFN-β application. A total of 16,978 probesets remained (FIGURE 2). Subsequent data filtering generated a list of 55 differentially expressed probesets at D2 versus baseline; respectively 285, 153 and 181 at later time points. Removal of unspecific Affymetrix probesets yielded the final DE probesets, as listed in ONLINE SUPPLEMENTARY TABLE 1 ([www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). GeneAnnot did not provide any information for Affymetrix control probesets

(e.g., AFFX-HUMISGF3A/M97935\_MB\_at binding STAT1) and several others (e.g., 223501\_at binding to TNFSF13B).

Filtered probesets and corresponding genes for each time pair are displayed in **FIGURE 3**. Plotting the data for each time point versus baseline as a function of p-values and MAID-scores produced so-called volcano plots. The distributions depicted preferential upregulation throughout time, in particular 1 month after first rIFN- $\beta$  injection (**FIGURE 3 & ONLINE SUPPLEMENTARY TABLE 2**; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). **TABLE 2** provides the enumeration and listing of the identified DEGs. Altogether, 339 unique probesets representing 269 genes were considered to be significantly regulated, with the maximum of DEGs determined at M1 (n = 175) (**FIGURE 2, TABLE 2**). The most consistently modulated genes throughout the analyzed time (n = 19 out of 269 DEGs) were divided into 18 up- and one

downregulated genes (**TABLE 2**). As expected, the majority represent established ISGs with antiviral properties, such as influenza-virus resistant gene *MX1*, and members of the 2–5A synthetase family (*OAS-2/-3*), as well as *EIF2AK2*, *RSAD2*, *IFI-44/-44L*, *IFIT-1/-2/-3* and *ISG15*, whereas to date *CMPK2*, *FFAR2* and *FCER1A* have not been identified in this context. When constricted time windows were analyzed, that is, until 1 year into therapy (n = 21 of 225 DEGs), or, excluding the first time point, until 2 years PTI (n = 48 of 252 DEGs), and the respective pairs possible, the greater overlap of reappearing DEGs was seen at later time points. Whereas most genes maintained the directionality over time, *CLU* and *IL1R2* switched from category increase at D2 to decrease at M1, respectively (**TABLE 2**; lower section).

Complementary to intersections in time were the varying amounts of newly appearing genes per sampling, rendering them time specific. In



**Figure 3. Overview of curated and filtered probeset numbers at each time point as result of pairwise comparison to baseline.** Volcano plots provide visual summaries of test statistics for all measured probesets in rIFN- $\beta$ -1b treated relapsing remitting MS patients (n = 25). The MAID score deflection is shown on the x-axis. The vertical bars at 2 and -2 indicate the cutoff used for differential expression of the represented genes. On the y-axis is the negative base 10 logarithm of the paired t-test value. The horizontal bar indicates the chosen significance threshold. Probesets that are significant at the  $\alpha = 0.05$  level are shown in the lighter color. Probesets that have been filtered following all three criteria as described in methods are displayed in diamond shape, located in upper right (increased expression) and upper left squares (decreased expression). The ratios compare the number of up- and down-regulated genes. \*See **ONLINE SUPPLEMENTARY TABLE 1** ([www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). \*\*see **ONLINE SUPPLEMENTARY TABLE 2** ([www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)).

DE: Differential expression; MAID: MA-plot-based signal intensity-dependent fold-change criterion.

## RESEARCH ARTICLE

Goertsches, Hecker, Koczan *et al.***Table 2. Temporal intersection of number of differentially expressed genes and respective change direction (increase or decrease).**

Time	Genes (n)	Complete time line	Three consecutive time points			Two time points		
			2 days/ 1 month/ 12 months/ 24 months	2 days/ 1 month/ 12 months/ 24 month	1 month/ 12 months/ 24 months	2 days	1 months	12 months
			269 genes	225 genes	254 genes	42 genes	175 genes	103 genes
2 days	42	19 (18,1)*	21 (20,1)†					
1 month	175	19 (18,1)*	21 (20,1)†	48 (46,2)§	24 (21,3)¶			
12 months	103	19 (18,1)*	21 (20,1)†	48 (46,2)§	23 (22,1)	69 (64,5)‡		
24 months	108	19 (18,1)*		48 (46,2)§	21 (20,1)	55 (52,3)	56 (52,4)**	
<b>Total</b>	<b>428</b>							

Consistently IFN- $\beta$ -1b subcutaneous responding genes throughout analyzed time ( $n = 19$  of 269) were divided in 18 up- and one down-regulated genes. 48 differentially expressed gene (DEGs) of M1/12/24 are reflected in 69 identically regulated DEGs between M1 and M12, and 56 DEGs when intersecting M12 and M24.

Genes sorted by change direction and alphabetically:

\*19 (18,1):

Increase = APOBEC3A, CMPK2, EIF2AK2, EPSTI1, FFAR2, HERC5, IFI44, IFI44L, IFIT1, IFIT2, IFIT3, ISG15, LY6E, MX1, OAS2, OAS3, RSAD2, SIGLEC1  
Decrease = FCER1A

Lists †,§ include additional 19 DEGs from list \*:  
†21 (20,1);  
Increase = IFI6, ZCCHC2  
Decrease =  
§48 (46,2);  
Increase = BAFF, CCR1, DDX58, DDX60, DDX60L, FBXO6, HERC6, IFI16, IFI27, IFI35, IRF7, ISG20, MARCKS, MS4A4A (*alias* CD20), MX2, OAS1, OASL, PARP14, PARP9, PLSCR1, PRIC285, RNF213, SAMD9, SAMD9L, SCO2, SP110, TNFSF10, XAF1  
Decrease = ITGA2B

Lists ¶,‡,\*\* include additional 21 DEGs from list † and additional 48 DEGs from list §, respectively:  
¶24 (21,3);  
Increase = KLRF1  
Decrease = CLU, IL1R2  
‡69 (64,5);  
Increase = C3AR1, CHMP5, GBP1, HPSE, IFI6, IFIH1, IFIT5, LGALS3BP, LOC26010, MAFB, NCOA7, SCO2, STAT1, TRIM22, TYMP, UBE2L6, ZCCHC2  
Decrease = C12orf39, ELOVL7, ITGB3  
\*\* 56 (52,4);  
Increase = ACSL1, ANKRD22, FPR2, LILRB2, MXD1, SAT1  
Decrease = GZMK, HOPX

comparison with D2, 169 (111 up, 58 down) further genes were filtered at M1, and 140 (74 up, 66 down) and 99 (60 up, 39 down) between subsequent time points, respectively (ONLINE SUPPLEMENTARY TABLE 3; www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152.).

To further describe the dynamics of the 19 consistently regulated genes, FIGURE 4 depicts time courses as a function of the MAID-score and the expression value, respectively (see values in ONLINE SUPPLEMENTARY TABLE 4; www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152). *FCER1A* was the exclusive gene to be down-regulated throughout time and the accentuated peak of the transcriptional effect at M1 PTI became apparent. ONLINE SUPPLEMENTARY FIGURE 1 (www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152) confirms the maximal change within the 25 individuals through the illustration of change calls per gene.

The patterns of gene expression revealed by real-time PCR were similar to those obtained from microarrays 1 month into therapy versus baseline. The mRNA measurements of both

techniques correlated significantly for 14 out of 15 genes. Details on real-time PCR experiments and respective correlations are presented in ONLINE SUPPLEMENTARY MATERIAL 1 (www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152).

### Gene ontology analysis

To functionally interpret the set of 269 selected genes, eight gene lists were tested for over-represented GO terms, stratified for each time pair for up- ( $n = 173$ ) and down-regulation ( $n = 96$ ) (TABLE 3, BOX 1 & ONLINE SUPPLEMENTARY TABLE 5; www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152). The significantly over-represented GO categories containing exclusively upregulated DEGs denote responses to viruses/other organisms/biotic stimuli, multiorganism processes (TABLE 3 & Box 1), homeostatic processes, immune effector processes, innate immune responses, (chemo)taxis, and regulation of programmed cell death, the latter being complemented by four additional genes (*EIF2AK2*, *GZMB*, *PRF1* and *RHOB*) in programmed cell death (ONLINE SUPPLEMENTARY TABLE 5; www.futuremedicine.com/

doi/suppl/10.2217/pgs.09.152). At D2 PTI, leukocyte-mediated immunity and humoral immune response, represented by *C1QA*, *C1QB* and *CLU*, were found to be over-represented, as were JAK-STAT cascade (*CCL2*, *NMI*, *STAT1*) and cytosis (*GZMB*, *PRFI*) at M1 PTI.

In comparison with the overall pronounced induction of genes, there was a smaller number of downregulated DEGs, which, in turn, were manifold appointed to GO terms of miscellaneous functions, including *IL-3* production, serotonin secretion and metabolic and biosynthetic processing of lipids, icosanoids, leukotrienes and alkenes earlier in therapy (ONLINE SUPPLEMENTARY TABLE 5; www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152). At M1 PTI cell adhesion, antiapoptosis, blood coagulation (*FCER1A* and *IL-8*), and at M12 PTI, cell surface

receptor-linked signal transduction (*ITGA2B*, *ITGB3*, *KLRB1*, *KLRG1* and *TGFBR3*) were significantly enriched in the gene set.

Up- and down-regulated genes were associated with 11 common GO terms. These categories belong mainly to immune defense programs, for example, response to stimulus, immune response, and defense response were found in upregulated DEGs throughout time (TABLE 3 & BOX 1), but only until one year for downregulated DEGs (ONLINE SUPPLEMENTARY TABLE 5; www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152).

#### ■ Construction of gene networks

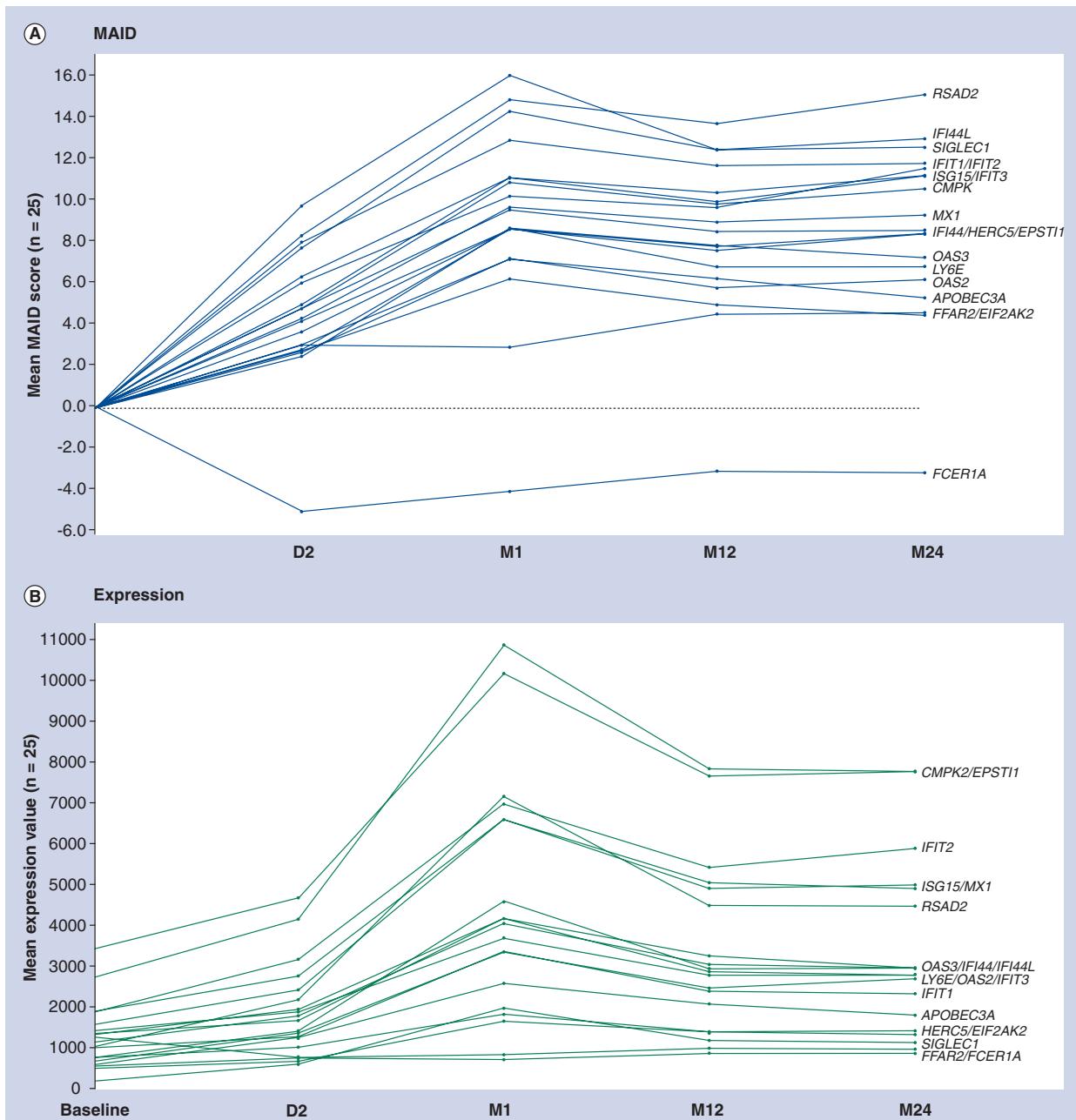
A total of 269 DEGs were inspected for evidence of 'direct interactions', applying the PA biological database. The resulting gene networks are

**Table 3. Biological process.**

Biological process	Gene size	Time	DEG	Count	ExpCount	Ratio	p-value
Response to stimulus (A; Box 1)	2109	2 D	38	17	5.96	4.6	$1.6 \times 10^{-5}$
		1 M	117	44	15.51	4.6	$4.7 \times 10^{-12}$
		12 M	85	40	12.10	6.4	$6.7 \times 10^{-14}$
		24 M	79	29	10.91	4.1	$1.2 \times 10^{-7}$
Immune system process (B; ONLINE SUPPLEMENTARY TABLE 5)	752	2 D		9	2.13	5.4	$1.8 \times 10^{-4}$
		1 M		24	5.53	5.7	$5.5 \times 10^{-10}$
		12 M		23	4.31	7.6	$1.2 \times 10^{-11}$
		24 M		15	3.89	4.8	$4.8 \times 10^{-6}$
Immune response (B; ONLINE SUPPLEMENTARY TABLE 5)	571	2 D		8	1.61	6.2	$1.5 \times 10^{-4}$
		1 M		22	4.20	6.8	$9.7 \times 10^{-11}$
		12 M		21	3.28	9.0	$3.9 \times 10^{-12}$
		24 M		14	2.95	5.9	$9.6 \times 10^{-7}$
Defense response (C; ONLINE SUPPLEMENTARY TABLE 5)	516	2 D		9	1.46	8.1	$9.5 \times 10^{-6}$
		1 M		21	3.79	7.1	$1.0 \times 10^{-10}$
		12 M		16	2.96	6.9	$2.5 \times 10^{-8}$
		24 M		9	2.67	3.8	$1.3 \times 10^{-3}$
Multiorganism process (D; ONLINE SUPPLEMENTARY TABLE 5)	267	2 D		5	0.76	7.7	$8.6 \times 10^{-4}$
		1 M		18	1.96	11.9	$6.7 \times 10^{-13}$
		12 M		17	1.53	15.2	$1.0 \times 10^{-13}$
		24 M		15	1.38	14.7	$4.3 \times 10^{-12}$
Response to biotic stimulus (D; ONLINE SUPPLEMENTARY TABLE 5)	236	2 D		5	0.67	8.7	$4.9 \times 10^{-4}$
		1 M		19	1.74	14.7	$5.5 \times 10^{-15}$
		12 M		16	1.35	16.0	$2.3 \times 10^{-13}$
		24 M		15	1.22	16.7	$7.1 \times 10^{-13}$
Response to other organism (D; ONLINE SUPPLEMENTARY TABLE 5)	167	2 D		5	0.47	12.5	$9.9 \times 10^{-5}$
		1 M		17	1.23	18.6	$3.5 \times 10^{-15}$
		12 M		16	0.96	23.4	$9.8 \times 10^{-16}$
		24 M		15	0.86	24.5	$4.3 \times 10^{-15}$
Response to virus (D; ONLINE SUPPLEMENTARY TABLE 5)	90	2 D		5	0.25	24.0	$5.0 \times 10^{-6}$
		1 M		17	0.66	38.4	$6.8 \times 10^{-20}$
		12 M		16	0.52	48.1	$3.5 \times 10^{-20}$
		24 M		15	0.47	50.0	$3.0 \times 10^{-19}$

Significantly over-represented gene ontology terms and respective upregulated DEGs at four time points versus baseline. Column by column, the table first lists the biological process term, for example, 'response to stimulus', its corresponding gene size ( $n = 2109$ ) of the contrasted gene universe ( $n = 12,377$ ), the time post-therapy initiation (day 2) and respective number of regulated DEGs ( $n = 38$ ), of which 17 genes (count) belong to 'response to stimulus'. With an expected count of 5.96, this led to an odds ratio of 4.6 and a p-value of  $1.6E-05$ . At the lower end, category 'response to virus' contained 90 related gene symbols, of which 5, 17, 16 or 15 appeared in respective upregulated DEG list, yielding a maximal odds ratio of 50 and p-value of  $3.5E-20$ , respectively. A total of 11 gene ontology terms appeared in up- and down-regulated DEG lists containing functionally related elements (see ONLINE SUPPLEMENTARY TABLE 5).

D: Day; DEG: Differentially expressed gene; M: Month.

RESEARCH ARTICLE Goertsches, Hecker, Koczan *et al.*


**Figure 4.** Time courses of 19 genes consistently modulated over 2 years in response to subcutaneous IFN- $\beta$ -1b as function of MAID score (A) and mean expression (B), respectively. Order of gene symbols in (A): if similar values at M24, gene symbol sequences correspond to higher values at preceding time point, that is, M12 or M1. See ONLINE SUPPLEMENTARY TABLE 4 ([www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)) for a more comprehensive data listing.

D: Day; MAID: MA-plot-based signal intensity-dependent fold-change criterion; M: Month.

shown in FIGURE 5 and ONLINE SUPPLEMENTARY FIGURE 2; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152). To survey their dynamics throughout the analyzed time course, time courses of genes are depicted as functions of the MAID score and expression value.

Taken together, 54 of 269 DEGs were connected at least once in networks, 15 genes of which appeared repeatedly (highlighted in FIGURE 5 and ONLINE SUPPLEMENTARY FIGURE 2; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)), mainly in the networks of DEGs

at M1 and M12. In total, 84 molecular interactions were attained. Of the 42 DEGs after 2 days into treatment, seven genes were connected (ONLINE SUPPLEMENTARY FIGURE 2; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). Of the 175 DEGs 1 month into therapy (FIGURE 5), 27 were arranged in two larger networks with overall stimulatory interactions in one network (FIGURE 5A), and functional linkage to mainly downregulated genes in the second (FIGURE 5B). The remaining 15 genes build an interacting trio (FIGURE 5C) and six pairs (FIGURE 5D–I). Out of the 103 genes modulated after 1 year of therapy, 14 were linked, and 2 years PTI 11 of 108 DEGs yielded two networks of five and four nodes, and a further gene pair (ONLINE SUPPLEMENTARY FIGURE 2; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)).

## Discussion

We report on genome-wide longitudinal gene expression changes in response to rIFN- $\beta$ -1b subcutaneous administration and present candidate genes for molecular response that have been identified in the PBMCs of 25 patients with RRMS. Differential gene expression was assessed using DNA microarrays and a combination of filtering criteria. As observed by other investigators using PBMCs from patients with MS diseases [12–28], our analyses showed the expression of multiple genes confirming the presence of IFN- $\beta$ . It has been shown that changes in mRNA expression can be seen within 2–4 h after drug administration [13,24,26,41]. However, our study was intended to grasp the ‘latest possible’ window of IFN-modulated genes, representing the high-hanging fruits of rIFN- $\beta$ -1b response. In Gilli *et al.*, *MX1* gene expression at 24 h after therapy initiation was 6.5-times higher in 19 IFN- $\beta$ -1b treated patients [41] and at 42 h after drug injection, Reder *et al.* reported modest but statistically significant upregulation of antiviral response genes in nine MS patients [24]. In conjunction with quantitative real-time PCR validation of presented microarray findings (see ONLINE SUPPLEMENTARY MATERIAL 1; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)), the authors’ confidence of genuine detection of regulated genes at 48 h is solid.

In order to account for cross-hybridizing Affymetrix probes [36], differentially expressed ‘promiscuous’ probesets were discharged based on information contained in the GeneAnnot database [37]. This lack of probeset specificity was neglected by preceding MS transcriptomics studies. Hence, several DEGs that may basically

suit MS and IFN- $\beta$  questions were reported, which are offered here for reconsideration (ONLINE SUPPLEMENTARY TABLE 1; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). Of note was a metallothionein MT2A probeset, cross-hybridizing with 18 additional genes, which was reported manifold as being differentially expressed [23–25,29], along with *MTIH* [25] and *MTIX* [24,25]. Furthermore, previously communicated probesets represent the proteasome activator (PSME2) [23], the ubiquitin specific peptidase (USP18) [22,23] and leucine aminopeptidase (LAP3) [9,22,24]. In contrast to others [24,29], dubious Affymetrix control probesets were eliminated. After discarding cross-hybridizing probes, one might lack genes that belong to functionally related and highly conserved multigene families, such as ribosomal proteins or immunoglobulins. However, occasionally, ‘promiscuous’ probesets can be superseded by a more specific alternative probeset for the same gene, thereby excluding the risk of cross-hybridization with any family member, whether they are presumed functionally identical or not. In the analyzed dataset, this was the case for the TNFSF13B, GBP1 and HBB probesets (ONLINE SUPPLEMENTARY TABLE 1; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)).

We ultimately filtered 269 DEGs, and thereof a subset of 96 previously reported ISGs [10–26,29] and 173 genes that have not been recognized to date in MS expression-array research with similar study design. Upregulated probesets representing type I IFN signaling molecules appeared in filtered lists, including RNA helicases DDX58 and IFIH1, IFN regulatory factors IRF7, IRF9, NMI, STAT1

### Box 1. Gene ontology category ‘response to stimulus’. A (n=61 increased differentially expressed genes.

#### 2 days

- C1QA, C1QB, CD163, CLU, EIF2AK2, IFI44, IFI6, IL1R2, ISG15, LY6E, MX1, OAS2, OAS3, PTGDS, RSAD2, SIGLEC1, TNFAIP6

#### 1 month

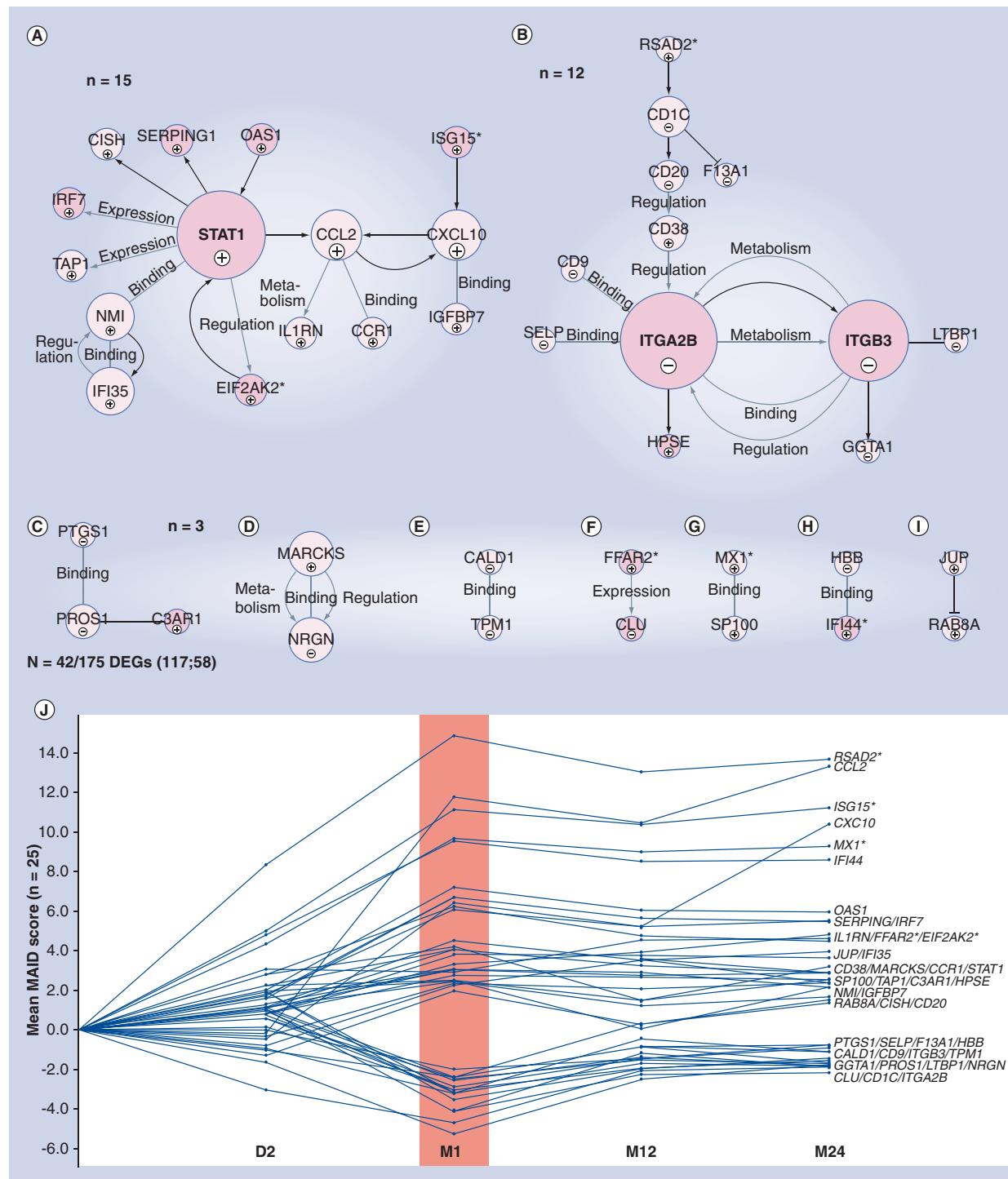
- BAFF, BST2, C3AR1, CCL2, CCR1, CXCL10, CYSLTR1, DDX58, EIF2AK2, GBP1, HPSE, IFI16, IFI35, IFI44, IFI6, IFIH1, IFITM1, IL1RN, IRF7, ISG15, ISG20, LGALS3BP, LY6E, MX1, MX2, NMI, OAS1, OAS2, OAS3, OASL, PLSCR1, PRF1, RSAD2, RTP4, SERPING1, SIGLEC1, STAT1, TAP1, TLR7, TNFSF10, TOR1B

#### 12 months

- ADM, AQP9, BAFF, C3AR1, CCR1, CLEC4E, DDX58, EIF2AK2, FCAR, FPR1, GBP1, HPSE, IFI16, IFI35, IFI44, IFI6, IFIH1, IL1RN, IRF7, ISG15, ISG20, LGALS3BP, LILRB2, LY6E, MX1, MX2, NFL3, OAS1, OAS2, OAS3, OASL, PLSCR1, SAD2, SERPING1, SIGLEC1, STAT1, TNFAIP6, TNFSF10, TRIM22, TYMP

#### 24 months

- BAFF, BST2, CCR1, CYSLTR1, DDX58, EIF2AK2, FAS, IFI16, IFI35, IFI44, IRF7, IRF9, ISG15, ISG20, LILRB2, LY6E, MX1, MX2, OAS1, OAS2, OAS3, OASL, PLSCR1, RSAD2, RTP4, SIGLEC1, STAT2, TNFSF10, TREM1

**RESEARCH ARTICLE** Goertsches, Hecker, Koczan *et al.*


**Figure 5. IFN- $\beta$ -1b subcutaneous specific gene networks derived from 175 differentially expressed genes and literature database information: baseline versus 1 month post-therapy initiation. See facing page.**

and STAT2 [7,8], as well as lymphocyte activation marker IFITM1. Within the first month of treatment, RNAs for genes implicated in

antiviral response (*MX-1* -2, *OAS-1* -2/-3, *GPBP1*, *EIF2AK2/PKR*, *RSAD2*, *ISG15*, *ISG20* and *PLSCR1*) were induced (TABLE 2).

**Figure 5. IFN- $\beta$ -1b subcutaneous specific gene networks derived from 175 differentially expressed genes and literature database information: baseline versus 1 month post-therapy initiation.** Upper-part IFN- $\beta$ -1b subcutaneous specific gene interaction networks derived for differentially expressed genes using Pathway Architect database information. Numeration in the lower left corner depicts the number of genes (differentially expressed genes) that were integrated in the network construction (N), stratified for up- and down-regulation in brackets. In the network plots, upregulated genes are indicated by a plus symbol and downregulated genes by minus. The degree of connectivity is visualized through the proportional node size, thereby stressing the relevance of selected differentially expressed genes. Important directed attributes were those of positive or negative regulation, that is, stimulation and inhibition (black arrows). Remaining interactions are presented by gray arrows and accompanying attribute labels. Genes appearing repeatedly over time are highlighted in gray. Gene symbols with an attached asterisk belong to the group of 19 genes with sustained expression changes over 2 years. Lower-part time courses of network relevant IFN- $\beta$ -1b subcutaneous responding genes for 2 years as function of the MAID score. Gene symbols with an attached asterisk belong to group of 19 consistently regulated genes over 2 years. The according time window of network construction is highlighted.  
MAID: MA-plot-based signal intensity-dependent fold-change criterion.

The experimental design of our analysis complements existing chip studies by describing expression dynamics over 2 years. In contrast to the small number of attained DEGs at day 2 PTI ( $n = 42$ ), an early window of therapy thoroughly investigated by others [13,23,24], the most intense transcriptional effect was seen at 1 month PTI ( $n = 175$ ). A comparable range of up to 3 or 6 months was covered by a few groups [17,18,21]. DEGs at year 1 ( $n = 103$ ) and 2 ( $n = 108$ ) described a relative steady state, and both revealed new genes in addition to those seen at earlier time points.

Of the 19 consistently modulated genes (TABLE 2), all genes but mitochondrial cytidine monophosphate kinase 2 (*CMPK2*), free fatty acid receptor 2 (*FFAR2*) and Fc fragment of IgE (*FCER1A*) were previously reported in other IFN- $\beta$ -related studies. Notably, still after 2 years of treatment, the 16 established genes are sustained indicators for biological IFN- $\beta$  responsiveness. Therefore, unrecognized upregulated *CMPK2* and *FFAR2* and downregulated *FCER1A* enlarge this group. The identification of the latter is promising in the sense that the associated  $\gamma$ -chain of the IgE receptor was found to be differentially expressed by other MS groups using Ficoll [19] or PaxGene [23], but the fact that it also forms subunits with alternative Fc receptors complicated its functional assignment. Mitochondrial membrane-bound proteins act as specific signaling adaptors, and the identified kinase *CMPK2* is a component of the salvage pathway for nucleotide synthesis. Interestingly, two participating enzymes of the pathway were recently detected by other MS researchers, namely deoxynucleotidase-2 (NT5M) [22] and deoxyguanosine kinase (DGUOK) [23]. The plasma membrane-bound short chain fatty acid receptor (*FFAR2*) has not been mentioned in context with IFN- $\beta$  or MS yet, but its contribution to the mechanisms of signal transmission through therapeutic drugs has received attention lately [42,43].

To learn more about the biological function of ISGs, GO term analysis provides a useful method to detect over-represented functional categories. Enriched GO annotations of DEGs revealed dual roles (up- and down-regulation) in mainly immune, inflammatory and defense responses (TABLE 3, BOX 1 & ONLINE SUPPLEMENTARY TABLE 5; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). Diverse categories delivered typical upregulated constituents of inflammatory cascades, such as the complement system (C1QA, C1QB and C3AR1), chemotactic signaling (CCL2, CCR1, CXCL10, FPR1 and TYMP), cell signaling (NMI) and cell binding (CLU). As expected, the category 'response to virus' was over-represented among exclusively upregulated genes, containing interferon-induced genes, interferon regulatory factors, Toll-like receptors and others (TABLE 3 & BOX 1). The same applied to the 'homeostatic process', disclosing the promising candidate gene *AQP9* [22] and the 'innate immune response'. The IFN- $\beta$  specific upregulation of TLR7, a cytoplasmatic receptor that recognizes ssRNA, might be an important factor to control the disease and has been discussed as therapeutic target elsewhere [44,45]. Accordingly, other members of the Toll-like receptor family were repeatedly detected in MS array research, such as increased TLR1 [22,23], TLR3 [19] and TLR5 [11,18]. Moreover, with regard to the known action of IFN- $\beta$ , we noticed potentially concurring features of DEGs that were assigned to GO terms of opposed function. One illustrative example was that 12 apoptosis regulator genes were induced, while four apoptosis inhibitors (ALOX12, CLU, PROK2 and SNCA) were suppressed (ONLINE SUPPLEMENTARY TABLE 5; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). Similarly, six upregulated genes belonging to 'chemotaxis/taxis' were functionally complemented to downregulated IL-8 in 'regulation of chemotaxis'.

Eventually, GO term analyses provided data-based evidence of interaction between determined DEGs, thus requiring the itemization

## RESEARCH ARTICLE

Goertsches, Hecker, Koczan *et al.*

of term constituents and their interconnection as nodes for IFN- $\beta$  specific gene networks. The gene interaction networks obtained from PA are of a rather phenomenological nature, as the provided interactions also include indirect relationships conveyed through intermediary molecules we do not know yet. Nonetheless, attained genes and respective interaction assignments presented in **FIGURE 5** and **ONLINE SUPPLEMENTARY FIGURE 2** ([www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)) provide new insights into IFN- $\beta$  mechanism of action. They confirm previously reported genes with identical or changed accessory targeting or regulating genes, and propose genes that have not been recognized in MS therapy chip research so far. For instance, three genes of the network of DEGs at D2 reappeared at M1: 'FFAR2 influences expression of *CLU*' and *IFI44* [24] (**ONLINE SUPPLEMENTARY FIGURE 2**; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)), but the latter was initially regulated by *IL8* [12] and then bound by *HBB* (**FIGURE 5H**). Further interactions of known and unknown DEGs during the first month of therapy were '*C1QB* binds *C1QB*', '*PTGS1* [19] binds *PROS1* that inhibits *C3AR1* [22]', '*MARCKS* [23] interacts with *NRGN*', '*TPM1* stimulates *CALD1*', '*MX1* [10] binds *SP100* [19]', and '*JUP* [12] inhibits *RAB8A*'.

The largest gene network (**FIGURE 5A**) comprises 15 upregulated genes. Of these, 13 genes had been previously detected in MS therapy research – *IRF7*, *ISG15*, *TAPI* [11], *CXCL10*, *IL1RN* [10], *STAT1*, *OASI* [12], *IFI35*, *NMI* [23], *CCR1*, *SERPING1* [22], *CCL2* [18] *EIF2AK2* [13] – whereas *CISH* (alias *SOCS1*) and *IGFBP7* were novel. By contrast, another gene network denoted nine downregulated versus three upregulated genes (**FIGURE 5B**). At least four genes were acknowledged: *CD38* [25], *CD9* [19], *RSAD2* [24] and *SELP* [10]. Here, *ITGA2B* and *ITGB3* were highly connected, as well as downregulated CD1c acting as stimulator for downregulated B lymphocyte cell surface antigen CD20.

At 1 year, *FPR1* [22] was connected to three other known MS-related genes: *C3AR1* [22], *GNLY* [23] and *FPR2* [24] (**ONLINE SUPPLEMENTARY FIGURE 2**; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). Here, *C3AR1* emerged as being functionally different to the members of a preceding network (**FIGURE 5C**). The interaction of upregulated B cell-activating cytokine *BAFF* [23,26] with fatty acid-degrading ligase *ACSL1* in the context of MS and IFN- $\beta$  expression analysis was yet unknown.

**ONLINE SUPPLEMENTARY FIGURE 2** ([www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)) also revealed regulatory interactions between four prominent MS related genes at 2 years PTI, displaying *FAS* [10] as a regulator of *XAF1* [24] and a stimulator of *TRAIL* [12], the latter being inhibited by *TNFRSF10C*, a decoy receptor of this cytokine. The significance of downregulated *MBP* may be both disease and drug related. The second network contained four established members of the IFN- $\beta$  signaling cascade [28], illustrating the regulatory action of *IRF9* [18], which forms a trimer with *STAT1* and -2 [19]. It acts upon interferon-responsive genes *IFIT2* [25] and *ISG15* [11], the latter being also described as a stimulator of *CXCL10* at M1 PTI (**FIGURE 5A**). Notably, the crucial function of *CXCL10* in an autoimmune disease such as Type I diabetes was recently supported [46].

Within the extracted gene networks, feedback and redundancy mechanisms were noticed. Regulatory feedback loops can be positive (reinforcing) or negative (self-balancing), and redundant links allow genes to sustain their effects on others even if some malfunction occurs. An exemplary positive feedback loop would be the regulatory chain '*CXCL10*→*CCL2*→*CXCL10*' (**FIGURE 5A**), and redundancy of interaction was evident between '*IFI44*–*IL8*–*MMP9*', '*ITGA2B*–*ITGB3*' and '*IRF9*–*STAT2*'. Such findings support the understanding of the IFN- $\beta$  mechanism of action, but intermediary steps that were not grasped with the applied transcriptomics approach still cause missing links in displayed gene networks, for example, the pleiotropic effects of a cytokine such as *IL8* (**ONLINE SUPPLEMENTARY FIGURE 2**; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)).

Focusing further on the novelties of this study, we assume that administered rIFN- $\beta$  lowers the activity of integrin-mediated signaling pathways (**ONLINE SUPPLEMENTARY TABLE 5**; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)) through the influence on a regulatory feedback mechanism, with downregulated *ITGA2B* and *ITGB3* as the most evident backbone (**FIGURE 5B & ONLINE SUPPLEMENTARY FIGURE 2**; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)). The latter also forms the cell surface glycoprotein IIb/IIIa, which mediates platelet aggregation by functioning as receptor for fibrinogen. This receptor serves as a target for several drugs. In addition, we found evidence for a modulation of B-cell-mediated immunity: the mean expression of *CD20* gene was significantly decreased at M1 compared with baseline. In the network

(FIGURE 5B), this target of rituximab therapy was downregulated by CD1c, an antigen-presenting protein that binds self and nonself lipid antigens and presents them to T-cell receptors on natural killer T-cells. Additionally, BAFF was upregulated in accordance with the literature [47]. Extended time dynamic knowledge of the surface molecule CD20 will be of crucial interest when considering the combination with rIFN- $\beta$  and questioning what kind of effects to expect: synergistic, unaltered or counteractive.

Finally, it is important to note that neutralizing antibodies are a recognized phenomenon in some people receiving rIFN- $\beta$  treatment. The determination of such interfering antibodies directed against rIFN- $\beta$ -1b and their biological significance is based on the most investigated ISG in MS research, *MX1* [48–50]. In presented expression data, a subgroup of patients did not reveal consistent *MX1* upregulation (ONLINE

SUPPLEMENTARY MATERIAL 2; [www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152](http://www.futuremedicine.com/doi/suppl/10.2217/pgs.09.152)), which was possibly due to the presence of neutralizing antibodies. However, regarding the inherent heterogeneity when analyzing *ex vivo* biomaterial and expected diverse responses to a low-dosage drug, our applied filtering system accounted for this by tolerating potential biological interference or pharmacogenetic nonresponses. Hence, ISGs down the cascades are still determined among the larger group of biologically responding patients, and represents a valuable insight into the mechanism of action of IFN- $\beta$ .

The clinical outcome was not the object of the presented analysis, and remained blinded to the main investigators. It is expected that biologically and/or clinically nonresponding patients are reflected in the generated dataset by unchanged or opposed gene expression. However, the confidence in detecting genuine mechanism-of-action players stands, as the applied filtering system accounts for such a degree of biological nonresponse (affecting ~30% of patients) in that tolerates up to 49% of patients showing no changes. The question of clinical response to rIFN- $\beta$  has been addressed by several research groups [12,17,26,51], and will be approached by our group in the near future, that is, performing subgroup classification on the basis of mRNA levels and developing predictive models of clinical response to rIFN- $\beta$ . This clinical branch will demand an exhaustive analysis, including the validation of clinically useful molecular biomarkers in larger cohorts, but would clearly reach beyond the currently stated question of IFN- $\beta$ -1b mechanism of action in time.

## Conclusion

The aim of the presented work was to analyze the pharmacogenomic effects in response to rIFN- $\beta$ -1b in time. It showed that the field of IFN- $\beta$ -regulation deserves further exploration, and we described previously unrecognized genes (*CMPK2*, *FCER1A* and *EPST1*) and maximal interactions 1 month into treatment. At 1 and 2 years PTI, data of disease progression and drug effects might be intertwined, but the consistent differential expression of ISGs suggests that, in the majority of analyzed individuals, the medication still affects the system and no adaption has set in. While the biological role(s) of many selected genes has been captured so far, it is essential to examine their molecular interactions that lead to regulatory cascades and signaling pathways, which still represent a large field that lies idle in MS therapy research.

## Future perspective

In drawing the concept of personalized medicine closer to MS patients, it is necessary to combine interrelating research areas, for example, pharmacogenomics with pharmacogenetics, which has already received thorough attention [52,53]. In addition, in order to define additional ISG identities and establish their functional relevance to the modulation of disease development and/or progression, that is, the immunological and beneficial effects of the pleiotropic agent IFN- $\beta$ , the employment of further accessible systems biology approaches [4,5,54–56] will be needed.

## Financial & competing interests disclosure

*Bayer Health Care and Bayer Vital GmbH (Germany) co-funded this study. The authors are grateful to the individuals who participated in this study. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.*

*No writing assistance was utilized in the production of this manuscript.*

## Ethical conduct of research

*The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.*

RESEARCH ARTICLE Goertsches, Hecker, Koczan *et al.*

## Executive summary

- Over 2 years (at time points: 2 days, 1, 12 and 24 months post-therapy initiation), 44,928 gene specificities were analyzed *ex vivo* for recombinant IFN- $\beta$  (rIFN- $\beta$ ) mechanism of action in 25 German relapsing remitting multiple sclerosis patients.
- All patients were prescribed exclusively with rIFN- $\beta$ -1b subcutaneously for their first administration of immunotherapy.
- To discover genuine transcriptional effects, data curation (eliminating noninformative probesets), gene filtering (paired student's t-test ( $p < 0.05$ ), adjusted fold-changes (MAID score  $> 2$ ), and increase/decrease labels ( $\geq 12$  individuals) were carried out.
- To consider cross-hybridizing Affymetrix probes, 'promiscuous' probesets were discharged based on the information contained in the GeneAnnot database.
- Overall, 269 genes were defined as differentially expressed genes, some of which appeared repeatedly at analyzed time points (2 days = 42; 1 month = 175; 1 year = 103; 2 years = 108).
- Most consistent modulated genes throughout time ( $n = 19$  of 269 differentially expressed genes) were divided in 18 up- and one down-regulated genes. 16 genes were previously reported, whereas three were hitherto unknown genes of biological response: mitochondrial cytidine monophosphate kinase 2 (CMKP2), free fatty acid receptor 2 (FFAR2) and Fc fragment of IgE (FCER1A).
- Functional interpretation by means of over-represented gene ontology terms in up- ( $n = 173$ ) and down-regulated ( $n = 96$ ) genes revealed dual roles in mainly immune, inflammatory and defense responses.
- Network constructions for each time point depicted interactions between new genes found to be influenced by rIFN- $\beta$  and enlarged information about established markers (e.g., BAFF) of biological response to treatment.
- This study shows the benefit of using microarray technology in determining biological response genes to rIFN- $\beta$  therapy. It has generated novel information likely to be of importance in furthering our understanding of Type I interferon biology in multiple sclerosis.

## Bibliography

Papers of special note have been highlighted as:  
■ of considerable interest

- 1 Hauser SL, Oksenberg JR: The neurobiology of multiple sclerosis: genes, inflammation, and neurodegeneration. *Neuron* 52(1), 61–76 (2006).
- 2 Hemmer B, Hartung HP: Toward the development of rational therapies in multiple sclerosis: what is on the horizon? *Ann. Neurol.* 62(4), 314–326 (2007).
- 3 Wiendl H, Toyka KV, Rieckmann P, Gold R, Hartung HP, Hohlfeld R; Multiple Sclerosis Therapy Consensus Group (MSTCG): Basic and escalating immunomodulatory treatments in multiple sclerosis: current therapeutic recommendations. *J. Neurol.* 255(10), 1449–1463 (2008).
- 4 Villoslada P, Steinman L, Baranzini SE: Systems biology and its application to the understanding of neurological diseases. *Ann. Neurol.* 65(2), 124–139 (2009).
- Overview of accessible systems biology approaches in the field of neurological diseases, and describes the driving forces to complement classic reductionist approaches in the biomedical sciences.
- 5 Quintana FJ, Farez MF, Weiner HL: Systems biology approaches for the study of multiple sclerosis. *J. Cell. Mol. Med.* 12(4), 1087–1093 (2008).
- 6 Stark GR, Kerr IM, Williams BR, Silverman RH, Schreiber RD: How cells respond to interferons. *Annu. Rev. Biochem.* 67, 227–264 (1998).
- 7 Der SD, Zhou A, Williams BR, Silverman RH: Identification of genes differentially regulated by interferon  $\alpha$ ,  $\beta$ , or  $\gamma$  using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* 95(26), 15623–15628 (1998).
- 8 de Veer MJ, Holko M, Frevel M *et al.*: Functional classification of interferon-stimulated genes identified using microarrays. *J. Leukocyte Biol.* 69(6), 912–920 (2001).
- 9 Rani MR, Shrock J, Appachi S, Rudick RA, Williams BR, Ransohoff RM: Novel interferon- $\beta$ -induced gene expression in peripheral blood cells. *J. Leukoc. Biol.* 82(5), 1353–1360 (2007).
- 10 Wandinger KP, Sturzebecher CS, Bielekova B *et al.*: Complex immunomodulatory effects of interferon- $\beta$  in multiple sclerosis include the upregulation of T helper 1-associated marker genes. *Ann. Neurol.* 50(3), 349–357 (2001).
- 11 Koike F, Satoh J, Miyake S *et al.*: Microarray analysis identifies interferon  $\beta$ -regulated genes in multiple sclerosis. *J. Neuroimmunol.* 139(1–2), 109–118 (2003).
- 12 Sturzebecher S, Wandinger KP, Rosenwald A *et al.*: Expression profiling identifies responder and non-responder phenotypes to interferon- $\beta$  in multiple sclerosis. *Brain* 126(6), 1419–1429 (2003).
- 13 Weinstock-Guttman B, Badgett D, Patrick K *et al.*: Genomic effects of IFN- $\beta$  in multiple sclerosis patients. *J. Immunol.* 171(5), 2694–2702 (2003).
- 14 Achiron A, Gurevich M, Magalashvili D, Kishner I, Dolev M, Mandel M: Understanding autoimmune mechanisms in multiple sclerosis using gene expression microarrays: treatment effect and cytokine-related pathways. *Clin. Dev. Immunol.* 11(3–4), 299–305 (2004).
- 15 Iglesias AH, Camelo S, Hwang D, Villanueva R, Stephanopoulos G, Dangond F: Microarray detection of E2F pathway activation and other targets in multiple sclerosis peripheral blood mononuclear cells. *J. Neuroimmunol.* 150(1–2), 163–177 (2004).
- 16 Hong J, Zang YC, Hutton G, Rivera VM, Zhang JZ: Gene expression profiling of relevant biomarkers for treatment evaluation in multiple sclerosis. *J. Neuroimmunol.* 152(1–2), 126–139 (2004).
- 17 Baranzini SE, Mousavi P, Rio J *et al.*: Transcription-based prediction of response to IFN- $\beta$  using supervised computational methods. *PLoS Biol.* 3(1), E2 (2005).
- 18 Satoh J, Nakanishi M, Koike F *et al.*: T cell gene expression profiling identifies distinct subgroups of Japanese multiple sclerosis patients. *J. Neuroimmunol.* 174(1–2), 108–118 (2006).
- 19 Satoh J, Nanri Y, Tabunoki H, Yamamura T: Microarray analysis identifies a set of CXCR3 and CCR2 ligand chemokines as early IFN- $\beta$ -responsive genes in peripheral blood lymphocytes *in vitro*: an implication for IFN- $\beta$ -related adverse effects in multiple sclerosis. *BMC Neurol.* 19, 6–18 (2006).
- 20 Palacios R, Goni J, Martinez-Forero I *et al.*: A network analysis of the human T-cell activation gene network identifies JAGGED1 as a therapeutic target for autoimmune diseases. *PLoS ONE* 2(11), E1222 (2007).
- 21 Annibali V, Di Giovanni S, Cannoni S *et al.*: Gene expression profiles reveal homeostatic dynamics during interferon- $\beta$  therapy in multiple sclerosis. *Autoimmunity* 40(1), 16–22 (2007).
- 22 Singh MK, Scott TF, La Framboise WA, Hu FZ, Post JC, Ehrlich GD: Gene expression changes in peripheral blood mononuclear cells from multiple sclerosis patients undergoing  $\beta$ -interferon therapy. *J. Neurol. Sci.* 258(1–2), 52–59 (2007).
- 23 Fernald GH, Knott S, Pachner A *et al.*: Genome-wide network analysis reveals the global properties of IFN- $\beta$  immediate transcriptional effects in humans. *J. Immunol.* 178(8), 5076–5085 (2007).

- First study in the field that incorporated network analysis to explore gene regulation in response to recombinant IFN- $\beta$ . Effective demonstration of the combination of data- and knowledge-driven analysis.
- 24 Reder AT, Velichko S, Yamaguchi KD *et al.*: IFN- $\beta$ 1b induces transient and variable gene expression in relapsing-remitting multiple sclerosis patients independent of neutralizing antibodies or changes in IFN receptor RNA expression. *J. Interferon Cytokine Res.* 28(5), 317–331 (2008).
- 25 Hilpert J, Beekman JM, Schwenke S *et al.*: Biological response genes after single dose administration of interferon  $\beta$ -1b to healthy male volunteers. *J. Neuroimmunol.* 199(1–2), 115–125 (2008).
- 26 Weinstock-Guttman B, Bhasi K, Badgett D *et al.*: Genomic effects of once-weekly, intramuscular interferon- $\beta$ 1a treatment after the first dose and on chronic dosing: Relationships to 5-year clinical outcomes in multiple sclerosis patients. *J. Neuroimmunol.* 205(1–2), 113–125 (2008).
- 27 Goertsches RH, Hecker M, Zettl UK: Monitoring of multiple sclerosis immunotherapy: from single candidates to biomarker networks. *J. Neurol.* 225 (Suppl. 6), 48–57 (2008).
- 28 Borden EC, Sen GC, Uze G *et al.*: Interferons at age 50: past, current and future impact on biomedicine. *Nat. Rev. Drug Discov.* 6(12), 975–990 (2007).
- Accessible and highly exhaustive description of interferon modes of action, and discussion of its relevance.
- 29 van Baarsen LGM, Vossenber S, Tijssen M *et al.*: Pharmacogenomics of interferon- $\beta$  therapy in multiple sclerosis: baseline IFN signature determines pharmacological differences between patients. *PLoS One* 3(4), E1927 (2008).
- 30 Satoh J, Illes Z, Peterfalvi A, Tabunoki H, Rozsa C, Yamamura T: Aberrant transcriptional regulatory network in T cells of multiple sclerosis. *Neurosci. Lett.* 422(1), 30–33 (2007).
- 31 Gneiss C, Tripp P, Reicharteder F *et al.*: Differing immunogenic potentials of interferon  $\beta$  preparations in multiple sclerosis patients. *Mult. Scler.* 12(6), 731–737 (2006).
- 32 McDonald WI, Compston A, Edan G *et al.*: Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the Diagnosis of Multiple Sclerosis. *Ann. Neurol.* 50(1), 121–127 (2001).
- 33 Yang IV, Chen E, Hasseman JP *et al.*: Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.* 3(11), research0062 (2002).
- 34 Hecker M, Goertsches RH, Engelmann R, Thiesen HJ, Guthke R: Integrative modelling of transcriptional regulation in response to antirheumatic therapy. *BMC Bioinformatics* 10(1), 262 (2009).
- Bioinformatics study that presents an analytical tool that incorporates and adjusts for the variability in gene expression intensities. Respective MA-plot-based signal intensity-dependent fold-change criterion (MAID) filtering was applied in the present study.
- 35 Pepper SD, Saunders EK, Edwards LE, Wilson CL, Miller CJ: The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics* 30(8), 273 (2007).
- 36 Ferrari F, Bortoluzzi S, Coppe A *et al.*: Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics* 15(8), 446 (2007).
- 37 Chalifa-Caspi V, Yanai I, Ophir R *et al.*: GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes. *Bioinformatics* 20(9), 1457–1458 (2004).
- 38 Ashburner M, Ball CA, Blake JA *et al.*: The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25(1), 25–29 (2000).
- 39 Falcon S, Gentleman R: Using GOstats to test gene lists for GO term association. *Bioinformatics* 23(2), 257–258 (2007).
- 40 Shannon P, Markiel A, Ozier O *et al.*: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13(11), 2498–2504 (2003).
- 41 Gilli F, Marnetto F, Caldano M *et al.*: Biological responsiveness to first injections of interferon- $\beta$  in patients with multiple sclerosis. *J. Neuroimmunol.* 158(1–2), 195–203 (2005).
- 42 Brown AJ, Jupe S, Briscoe CP: A family of fatty acid binding receptors. *DNA Cell Biol.* 24(1), 54–61 (2005).
- 43 Hirasawa A, Hara T, Katsuma S, Adachi T, Tsujimoto G: Free fatty acid receptors and drug discovery. *Biol. Pharm. Bull.* 31(10), 1847–1851 (2008).
- 44 Hemmi H, Kaisho T, Takeuchi O *et al.*: Small anti-viral compounds activate immune cells via the TLR7 MyD88-dependent signaling pathway. *Nat. Immunol.* 3(2), 196–200 (2002).
- 45 O'Neill LA: Targeting signal transduction as a strategy to treat inflammatory diseases. *Nat. Rev. Drug Discov.* 5(7), 549–563 (2006).
- 46 Schulthess FT, Paroni F, Sauter NS *et al.*: CXCL10 impairs  $\beta$  cell function and viability in diabetes through TLR4 signaling. *Cell Metab.* 9(2), 125–139 (2009).
- 47 Krumbholz M, Faber H, Steinmeyer F *et al.*: Interferon- $\beta$  increases BAFF levels in multiple sclerosis: implications for B cell autoimmunity. *Brain* 131(6), 1455–1463 (2008).
- 48 Gilli F, Marnetto F, Caldano M *et al.*: Biological markers of interferon- $\beta$  therapy: comparison among interferon-stimulated genes *MxA*, *TRAIL* and *XAF-1*. *Mult. Scler.* 12(1), 47–57 (2006).
- 49 Hesse D, Sellebjerg F, Sorensen PS: Absence of *MxA* induction by interferon  $\beta$  in patients with MS reflects complete loss of bioactivity. *Neurology* 73(5), 372–377 (2009).
- 50 Sellebjerg F, Krakauer M, Hesse D *et al.*: Identification of new sensitive biomarkers for the *in vivo* response to interferon- $\beta$  treatment in multiple sclerosis using DNA-array evaluation. *Eur. J. Neurol.* (2009) (Epub ahead of print).
- 51 Comabella M, Lünemann JD, Río J *et al.*: A type I interferon signature in monocytes is associated with poor response to interferon- $\beta$  in multiple sclerosis. *Brain* 2(Pt 12), 3353–3365 (2009).
- 52 O'Doherty C, Villoslada P, Vandenbroucke K: Pharmacogenomics of Type I interferon therapy: a survey of response-modifying genes. *Cytokine Growth Factor Rev.* 18(3–4), 211–222 (2007).
- 53 Vandenbroucke K, Matute C: Pharmacogenomics of the response to IFN- $\beta$  in multiple sclerosis: ramifications from the first genome-wide screen. *Pharmacogenomics* 9(5), 639–645 (2008).
- 54 Hwang D, Rust AG, Ramsey S *et al.*: A data integration methodology for systems biology. *Proc. Natl Acad. Sci. USA* 102(48), 17296–17301 (2005).
- 55 Bauch A, Superti-Furga G: Charting protein complexes, signaling pathways, and networks in the immune system. *Immunol. Rev.* 210, 187–207 (2006).
- 56 Lamb J, Crawford ED, Peck D *et al.*: The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313(5795), 1929–1935 (2006).

### Websites

- 101 Pathway Architect 2.0.1 (Iobion) [www.iobion.com](http://www.iobion.com)
- 102 National Center for Biotechnology Information Gene Expression Omnibus [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)

## 6. Manuscript IV

---

### **Network analysis of transcriptional regulation in response to intramuscular interferon-beta-1a multiple sclerosis treatment**

Michael Hecker, Robert Hermann Goertsches, Christian Fatum, Dirk Koczan, Hans-Jürgen Thiesen, Reinhard Guthke, and Uwe Klaus Zettl

*Pharmacogenomics J.* submitted January 25, 2010.

---

## **Network analysis of transcriptional regulation in response to intramuscular interferon-beta-1a multiple sclerosis treatment**

Michael Hecker<sup>1</sup>, Robert Hermann Goertsches<sup>2,3</sup>, Christian Fatum<sup>3</sup>, Dirk Koczan<sup>2</sup>, Hans-Juergen Thiesen<sup>2</sup>, Reinhard Guthke<sup>1</sup>, Uwe Klaus Zettl<sup>3</sup>

<sup>1</sup> Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knoell-Institute,  
Beutenbergstr. 11a, D-07745 Jena, Germany,  
e-mail: {michael.hecker, reinhard.guthke}@hki-jena.de,  
phone: +49-3641-5321083, fax: +49-3641-5320803

<sup>2</sup> University of Rostock, Institute of Immunology, Schillingallee 70, D-18055 Rostock, Germany,  
e-mail: {dirk.koczan, hans-juergen.thiesen}@med.uni-rostock.de

<sup>3</sup> University of Rostock, Department of Neurology, Gehlsheimer Str. 20, D-18147 Rostock,  
Germany, e-mail: {robert.goertsches, christian.fatum, uwe.zettl}@med.uni-rostock.de

Running title: Transcriptional effects of IFN- $\beta$  therapy in MS

Funding: This study was supported by grants from the German Federal Ministry of Education  
and Research (BMBF, BioChancePlus, 0313692D), and partially funded by Biogen Idec.

## Abstract

Interferon- $\beta$  (IFN- $\beta$ ) is one of the major drugs for multiple sclerosis (MS) treatment. The purpose of this study was to characterize the transcriptional effects induced by intramuscular IFN- $\beta$ -1a therapy in patients with relapsing-remitting form of MS. By using Affymetrix DNA microarrays, we obtained genome-wide expression profiles of peripheral blood mononuclear cells of 24 MS patients within the first four weeks of IFN- $\beta$  administration. We identified 121 genes that were significantly up- or down-regulated, in particular one week after start of therapy. Eleven transcription factor binding sites (TFBS) are overrepresented in the regulatory regions of these genes, including those of IFN regulatory factors and NF- $\kappa$ B. We then applied TILAR, a novel integrative algorithm for deriving gene regulatory networks from gene expression data and TFBS information, to reconstruct the underlying network of molecular interactions. An NF- $\kappa$ B-centered sub-network of genes was higher expressed in patients with IFN- $\beta$ -related side effects. Expression changes were confirmed by real-time PCR and literature mining was applied to evaluate network inference accuracy.

**Keywords:** interferon-beta, multiple sclerosis, DNA microarray, network analysis

## Introduction

Multiple sclerosis (MS) is an idiopathic inflammatory disorder of the central nervous system and the most common disabling neurologic disease of young adults. It is a life-long disease that affects the nervous system by destroying the protective covering (myelin) that surrounds nerve fibers, thereby provoking impaired nerve conduction. Genetic susceptibility, environmental exposure and immune dysregulation are thought to play a significant role in its pathogenesis. An autoimmune basis is assumed to cause lesions in the brain and spinal cord, but it is less clear how the autoimmune processes originate and maintain [1-3]. At present there is no cure for MS, and existing therapies are designed primarily to prevent lesion formation and brain atrophy, decrease the rate and severity of relapses and delay the resulting disability by reducing levels of inflammation. Interferon- $\beta$  (IFN- $\beta$ ) is currently the most established treatment for controlling the exacerbations in relapsing-remitting MS and several studies have confirmed its clinical benefit [4-6].

IFN- $\beta$  is a natural human pleiotropic cytokine with antiproliferative and immunomodulatory activities that is produced by various cell types including fibroblasts and macrophages. It exerts its biological effects by binding to specific cell surface receptors. This binding initiates a cascade of intracellular events which leads to the activation of transcription factors (TFs) including STATs, IFN regulatory factors and NF- $\kappa$ B [7]. These translocate to the nucleus and drive the expression of numerous genes that could serve as biological markers of IFN- $\beta$  activity (e.g. MX1, TNFSF10, B2M and IFIT1). Transcript levels of these biological response markers increase within 12 hours of IFN- $\beta$  administration and remain elevated for at least 3 days [8-11]. While the initial steps of the intracellular signaling pathways that are triggered by IFN- $\beta$  have been delineated in great detail, the mechanisms instrumental for its therapeutic efficacy in MS are still poorly understood. It is likely that of the many (immune system) processes that are influenced by IFN- $\beta$  - directly or indirectly as a consequence of the activity of induced proteins - only a small proportion is responsible for the clinical benefit. This suggests a cascade of events following IFN- $\beta$  administration, in which some are beneficial, others may have no effect on MS, and still others may be deleterious and cause side effects. The balance of all these effects is favorable in most MS patients and leads to reduced disease activity and lowered accumulation of disease burden. However, clinical trials demonstrated that MS patients exhibit considerable interindividual heterogeneity in their clinical course and response to IFN- $\beta$  therapy. Approximately one-third of the patients suffers from a higher or identical annual relapse rate while on treatment than before (non-responders) [12]. At present, no

established markers capable of predicting either favorable or detrimental responses to IFN- $\beta$  are available.

Today, different IFN- $\beta$  medications are available for MS. Three have been approved as first-line therapies for the treatment of relapsing-remitting MS in the mid-1990s: Avonex (IFN- $\beta$ -1a intramuscular; Biogen Idec, Cambridge, MA, USA), Rebif (IFN- $\beta$ -1a subcutaneous; Merck Serono, Darmstadt, Germany) and Betaferon (IFN- $\beta$ -1b subcutaneous; Bayer Schering, Leverkusen, Germany). Intramuscular (i.m.) IFN- $\beta$ -1a is given once a week, while subcutaneous (s.c.) IFN- $\beta$ -1a and IFN- $\beta$ -1b require three to four injections per week. IFN- $\beta$ -1a i.m. appears to be well tolerated with 4% of treated patients discontinuing injections due to adverse events in the pivotal clinical trial [5]. Moreover, it has the lowest incidence (approximately 2%) of neutralizing antibodies [13].

It has been repeatedly demonstrated that IFN- $\beta$ -mediated gene regulatory effects can be accessed by expression profiling of peripheral blood cells using DNA microarrays [14,15]. In the recent past, a few high-throughput analyses have been completed in an attempt to explicitly characterize the global transcriptional changes in the blood that occur in response to IFN- $\beta$ -1a i.m. (table 1). In a pharmacogenomic study by Weinstock-Guttman et al. ~4000 genes were measured, of which about 1500 were identified as up- or down-regulated after the first dose and after chronic administration of IFN- $\beta$ -1a i.m. [10]. This showed that the therapy induces changes in the expression of many genes (e.g. cytokines and cell adhesion molecules), and that IFN- $\beta$  has effects on multiple processes.

However, existing studies primarily explored the immediate gene expression changes few hours after therapy onset, and to a lesser extent the (possibly indirect) regulatory effects of IFN- $\beta$  that appear at a late stage between two subsequent injections and sustain for a longer period of time. Moreover, most reports just list differentially expressed genes, but since genes tend to interact, it is more meaningful to arrange them in gene regulatory networks (GRNs) based on expression data and known molecular interactions. A GRN, in principle, denotes the assembly of regulatory effects that conduct the mutual transcriptional control of a set of genes. Various modeling approaches have been proposed to reconstruct GRNs from experimental data on the basis of different mathematical concepts and learning strategies, and distinct degrees of abstraction. Novel network inference algorithms integrate diverse types of data (e.g. gene expression data and protein-DNA interaction data), incorporate prior biological knowledge (e.g. from scientific literature) and use specific modeling constraints to obtain more accurate GRN models [18,19]. A network analysis could provide testable hypotheses about the drug's mechanisms of action and may have clinical

implications by accentuating gene sub-networks with expression differences between responders and partial responders or patients with and without side effects to therapy. In 2007, Fernald et al. published the first and only study that incorporated computational network inference to investigate gene regulation in response to IFN- $\beta$ -1a i.m. [17]. They used mutual information as a similarity measure to derive a large network of genes.

Here, we used DNA microarrays to measure the transcriptional profile of peripheral blood mononuclear cells (PBMC) from 24 MS patients within the first month of weekly intramuscular injection of IFN- $\beta$ -1a. The data allowed to assess sustained changes in expression and we analyzed the functional characteristics and TF binding sites (TFBS) of the genes up- or down-regulated during therapy. Using an integrative modeling approach we reconstructed a GRN in which TFs are linked to the modulated genes to provide a deeper molecular understanding of the underlying therapeutic mechanisms early in therapy.

## Materials and methods

### Study population

A total of 24 Caucasian patients (18 females / 6 males, mean age 35.8 years; table 2) diagnosed with relapsing-remitting MS by McDonald criteria [20] were analyzed in this study. The patients were prescribed a first therapy with 30 µg, once-weekly, intramuscular IFN- $\beta$ -1a (Avonex; Biogen Idec, Cambridge, MA, USA). None of the patients had previously been medicated with immunomodulatory or immunosuppressive agents or had ever received cytotoxic treatments, and all were free of glucocorticoid treatment for at least 30 days prior to blood extraction. Patients were assessed neurologically and rated using the Expanded Disability Status Scale (EDSS) at regular intervals. The study was approved by the University of Rostock's ethics committee and carried out according to the Declaration of Helsinki.

### Gene expression analysis using microarrays

With informed consent, 15 ml peripheral venous EDTA blood samples were taken from all patients immediately before first, second and fifth IFN- $\beta$  injection, i.e. at baseline as well as one and four weeks post therapy initiation. The samples were always collected at the same time of the day and processed within one hour. Total RNA of Ficoll-isolated PBMC from each sample was extracted following manufacturer's protocol (RNeasy; Qiagen, Hilden, Germany). We used PBMC instead of whole blood to reduce interferences of globin RNA and thus increase the sensitivity of the microarray hybridization results. Initial RNA and final cRNA concentrations were determined spectrophotometrically by a Nanodrop 1000 (Thermo Fisher Scientific, Waltham, MA, USA) and quality control was performed by native ethidium bromide agarose gel electrophoresis. Samples of 7 µg total RNA were labeled and hybridized to Affymetrix Human Genome U133 A and B arrays in accordance with the supplier's instructions. The arrays were scanned at 3 micron resolution using the Hewlett Packard GeneArray Scanner G2500A (Affymetrix, Santa Clara, CA, USA). The raw data were stored according to the MIAME standard and are available from Gene Expression Omnibus (accession number GSE19285, <http://www.ncbi.nlm.nih.gov/geo/>).

## Validation of the microarray data by real-time PCR

Quantitative real-time reverse transcription polymerase chain reaction (real-time PCR) was used to confirm changes in gene expression revealed by the DNA microarray experiments. We measured the expression levels of 15 selected genes in a subset of the samples. Details on the experimental procedure as well as on the analysis of the real-time PCR data are described in the supplemental document.

## Microarray data preprocessing

In the applied Affymetrix microarrays most probesets include probes matching transcripts from more than one gene and probes which do not match any transcribed sequence. Therefore, we utilized custom chip definition files (CDFs) that are based on the information contained in the GeneAnnot database version 1.8 [21] (<http://bioinfo2.weizmann.ac.il/geneannot/>). GeneAnnot-based CDFs are composed of probesets including exclusively probes matching a single gene and thus allow for a more reliable determination of transcript levels. We used version 1.5.0 of the custom CDFs (for HG-U133 A and B) and the MAS5.0 algorithm to preprocess the raw probe intensities. Data normalization was performed, separately for the arrays of type A and B, by a loess fit to the data with *span*=0.05 (using R package affy). Each A- and B-chip yielded mRNA abundances of 11220 and 6771 human genes, respectively. For the 2257 genes that were measured with both chip types, we used the signal intensities of the A-chip.

## Filtering differentially expressed genes

To identify genes substantially up- or down-regulated within the first month of therapy, we used a combination of two criteria: a test of statistical significance and a fold-change variant. First, we analyzed which genes show significant expression changes by use of a paired *t*-test comparing the baseline expression levels with the levels at one week and four weeks into therapy. To further narrow down the filtering result, we evaluated signal intensity-dependent fold-changes (MAID filtering) [22] (<http://www.hki-jena.de/index.php/0/2/490>). In this way, we took into account that the variability in the (log) fold-changes increases as the measured signal intensity decreases [23]. The absolute value of a gene's MAID-score is higher, the more its expression level is altered after start of therapy, and it is generally lower for weakly expressed genes. Finally, we selected the protein-coding genes having |MAID-score|>3.0 and *t*-test *P*-value<0.01 as IFN- $\beta$ -responsive genes.

To provide an estimate of the number of genes passing the filtering by chance, a permutation test was performed. The data set was permuted 1000 times by randomly rearranging the temporal sequence of the data of each patient. The same filtering criteria as described above were applied to each permutation.

In addition, we analyzed the genes with significant expression changes to identify a subset of genes whose expression correlates with IFN- $\beta$ -related side effects. For this purpose, we compared the baseline transcript levels of patients with and without side effects using a two-sided two-sample *t*-test with the significance level at  $\alpha=0.05$ .

### **Gene Ontology analysis**

We examined the set of IFN- $\beta$ -responsive genes for overrepresented Gene Ontology (GO) terms using GOstats, a software package written in R [24]. Each GO term was tested whether it is significantly associated to the list of filtered genes in comparison to the 15734 genes measured in total. The analysis was performed for gene functional annotations of the biological process GO category provided by the Bioconductor annotation package org.Hs.eg.db version 2.2.6 (<http://www.bioconductor.org>).

### **Transcription factor binding site analysis**

To reconstruct the regulatory interactions between the genes with expression changes in response to IFN- $\beta$ , we applied a GRN inference algorithm that integrates information on TFBS as prior knowledge. Evolutionarily conserved TFBS were derived from the tfbsConsSites track of the UCSC database build hg18 (<http://genome.ucsc.edu>). The track data were generated using position weight matrices (PWMs) of TFBS contained in the public Transfac database (version 7.0, <http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/>). For the whole human genome 3837187 TFBS predictions associated to 258 different PWMs can be retrieved from the tfbsConsSites track (as of November 19, 2009). With this at hand, we screened the regulatory region of each of the 15734 measured genes for TFBS occurrences. The regulatory region was specified as the genomic sequence 1000 bp up- and downstream of the transcription start site stated in the GeneCards database version 2.39 (<http://www.genecards.org>). Furthermore, to take into account the inherent redundancy of the Transfac database, we grouped very similar or identical sequence motifs by use of STAMP [25]. In this way, we could reduce the 258 Transfac PWMs to 101 distinct DNA-binding patterns. Using a hypergeometric test, we identified a subset of

(consolidated) TFBS motifs overrepresented in the regulatory regions of the genes at the significance level  $\alpha=0.1$ . This information corresponds to a list of predicted TF-gene interactions that can serve as a template for inferring the GRN. A TF-gene interaction represents a physical interaction, i.e. a TF (or a group of TFs with similar binding specificity) binds at least once the DNA at the region that encompasses the transcription start site of a gene and thus presumably regulates its transcription.

### **Integrative gene regulatory network modeling**

We applied the TFBS-integrating least angle regression (TILAR) algorithm to construct a GRN model of IFN- $\beta$ -responsive genes. TILAR is a novel method for inferring GRNs from gene expression data, incorporating known or predicted TFBS and, if available, literature mining information (adaptive TILAR) [22]. The modeling approach distinguishes two types of network nodes: genes (that were selected for identifying the regulatory interactions between them) and TFs (whose binding sites are overrepresented in the regulatory regions of the genes). The algorithm then assigns (directed) TF-gene and gene-TF interactions (network edges) by fitting a system of linear equations to the genes' expression levels. In comparison to TF-gene interactions, gene-TF interactions can have different meanings, e.g. the gene itself might encode a transcriptional regulator of the TF, or the gene product controls (possibly via a signaling cascade) the activity of the TF at the proteome level. Using both types of interactions, the modeling considers that genes usually regulate other genes indirectly through the activity of one or more TFs.

The actual GRN inference problem is formulated as a linear regression equation that satisfies specific constraints on the network structure. The model is constrained to be sparse reflecting that genes are regulated by a limited number of regulators. Moreover, TILAR includes only a subset of the putative TF-gene interactions obtained from TFBS analysis. As TILAR not necessarily employs all those TF-gene interactions, the method considers the fact that they are predicted and therefore not all of them might refer to biologically functional binding sites. The regression coefficients (i.e. the nominal parameters in the model) to be estimated by least angle regression [26] then specify the presence and strength (edge weights) of each possible gene-TF interaction. The final model was determined by 10-fold cross-validation (to avoid overfitting to the data), visualized with Cytoscape 2.6.0 [27], and tested for scale-freeness according to Clauset et al. [28]. For a more complete description of the modeling approach, we refer to our original paper on TILAR [22].

The major advantage of this inference technique is that only few model parameters are sufficient to define a complex network, which is still readily interpretable in terms of true molecular interactions. Moreover, TF expression levels are not required for the network reconstruction, since the activity of TFs is modeled implicitly. This is beneficial, as TF expression seldom correlates with TF activity. We evaluated and compared the performance of TILAR (including the adaptive variant) and five different GRN inference methods using literature mining information. A detailed report on this benchmarking analysis can be found in the supplemental document. We supply R codes for TILAR at our institute's website (<http://www.hki-jena.de/index.php/0/2/490>).

## Results

### Patient characteristics

The demographic and clinical characteristics of the 24 patients in our study are summarized in table 2. Twenty patients had systemic side effects such as flu-like symptoms (including fever, chills, asthenia, myalgia and headaches) within the first 3 months on treatment. Two patients experienced a relapse in this observation period.

### Transcriptional changes in response to IFN- $\beta$ therapy

The preprocessing of the microarray data resulted in an expression data set of 15734 different genes and 72 PBMC samples. By filtering for genes significantly up- or down-regulated after therapy onset, we detected gene expression changes that were common among the patients. We identified 102 genes as differentially expressed at week 1 versus baseline, and 24 genes at week 4. Altogether, 72 genes were found up-regulated and 49 genes down-regulated, comprising a set of 121 genes in total (supplemental table). Most of the filtered genes showed significant expression changes during the first week of IFN- $\beta$ -1a i.m. therapy (figure 1).

The permutation test disclosed that the number of 121 differentially expressed genes is significantly higher than would be expected by chance. When randomly shuffling the sampling time points for each patient for 1000 times and analyzing each permuted data set for modulated genes, between 8 and 89 genes were filtered (on average 21.4). Hence, the number of filtered genes was below 121 in each permutation, which implies an empirical *P*-value for the actual filtering result of  $<0.001$ . This shows that (most of) the identified mRNA changes are due to the therapy.

The overall patterns of gene expression revealed by real-time PCR were similar to those obtained using microarrays. The mRNA measurements of both techniques correlated significantly for all 15 remeasured genes. When comparing the expression levels before first and second IFN- $\beta$  injection, real-time PCR analysis also confirmed the significance of expression changes for all of these genes. Full results of the real-time PCR experiments are presented in the supplemental document.

### Functional analysis of IFN- $\beta$ -responsive genes

The result of the GO analysis (table 3) shows that most of the filtered genes are known to participate in immune system processes. Genes associated to "positive regulation of immune

response" are significantly enriched in the gene set. The corresponding GO category (GO:0050896) comprises 26 out of the 121 genes, including TNFSF10, TLR7, FCER1A, IL1R2, OAS1 and AQP9. These genes have quite diverse immune functions. OAS1, a member of the 2-5A synthetase family, is an essential IFN-inducible protein involved in the innate immune response to viral infection [29]. The water-selective membrane channel AQP9 is thought to play a role in the immunological function of leukocytes [30]. FCER1A is an Fc receptor for IgE molecules and has been demonstrated to induce NF-κB activation in monocytes [31]. Apart from that, 10 of the genes in the category "positive regulation of immune response" constitute a subgroup of genes annotated with the GO term "B cell mediated immunity" (GO:0042221).

G proteins (GNAZ and GNG8) and G protein-coupled receptors (GPR20, GPR44, GPR56 and GPR97) also showed significant expression changes in response to the therapy. Modifications in the activity of GPRs characterize lymphocytes from different chronic immune disorders including MS, and it is assumed that IFN-β-1a affects the expression of molecules responsible for GPR regulation [32]. Besides, we noted an up-regulation of specific adhesion molecules including the integrins (JAM3, ESAM, ITGA2B and ITGB3). Adhesion molecules are believed to regulate the transmigration of blood leukocytes across the blood-brain barrier (see discussion).

### **Comparison with other expression profiling studies on IFN-β-1a i.m.**

Several of the 121 filtered genes have already been described as genes altered at the transcript level in response to intramuscular IFN-β-1a therapy. In the study by Fernald et al., the genes CSF1R, CST7, OAS1 and SNCA were found modulated by the treatment, too [17]. Their study focused on the blood expression changes within two days after drug administration. CSF1R, which was down-regulated after one and four weeks into therapy in our study, is a receptor for colony stimulating factor 1, a cytokine that controls differentiation and function of monocytes and macrophages [33]. The down-regulation of CSF1R was confirmed by real-time PCR analysis (supplemental document). In another study, Singh et al. examined 5 MS patients before and 24 h after initial therapeutic dose of IFN-β-1a i.m. and identified a set of 132 differentially expressed genes [16]. At least 8 of these genes also occur in our filtering result: TNFSF10, OAS1, JUP, AQP9, FGFBP2, MS4A4A, TYMP and FAM26F. The mRNA levels of TNFSF10 are known to be increased in PBMC from MS patients compared to healthy controls [34] and even have been reported to be predictive of clinical responsiveness to IFN-β [35]. TNFSF10 encodes a cytokine of the tumor necrosis factor (TNF) ligand family that induces apoptosis and activation of NF-κB. Regulation of

TNFSF10 function takes place at the level of receptor expression, while decoy receptors such as TNFRSF10C can inhibit TNFSF10 from binding to TNF receptors capable of mediating apoptosis [36]. We found TNFRSF10C significantly up-regulated after first injection of IFN- $\beta$ -1a i.m., but the expression returned to baseline levels at the 4 weeks time point.

### **Identification of putative TF-gene interactions**

Genes responsive to IFN- $\beta$ -1a are under control of TFs, whose activities are (indirectly) affected by the drug. Therefore, we analyzed the genes' regulatory regions for occurrence of overrepresented TFBS. Conserved binding sites were found enriched for 11 consolidated TFs. These 11 TFs connect 77 out of the 121 genes through 152 TF-gene interactions, and each TF is linked to at least 8 genes (table 4). This information was used as a template for inferring the underlying GRN by TILAR.

It is important to note that some TFs with highly similar DNA-binding properties were grouped to one TF entity, e.g. the interferon regulatory factors (IRF) 1 and 2. IRF1 and IRF2 are structurally similar and bind to the same regulatory elements. However, they have distinct or even antagonistic functions. IRF2 is a repressor that competitively inhibits the IRF1-mediated transcriptional activation of interferons and IFN-inducible genes [37].

A subset of 21 genes possess at least one TFBS for NF- $\kappa$ B (NFKB1, NFKB2, REL, RELA). NF- $\kappa$ B proteins are key regulators in the transcription of many inflammatory genes and are activated by various intra- and extra-cellular stimuli, e.g. cytokines like IFN- $\beta$ . They can be found in numerous cell types that express cytokines, cell adhesion molecules and acute phase proteins.

### **Integrative network modeling**

On the basis of the gene expression data (121 genes, 72 samples) and the (predicted) TF-gene interactions, we constructed a GRN model using the TILAR algorithm. The modeling was constrained to use only a subset of the putative TF-gene interactions. Here, 102 out of the 152 predicted TF-gene interactions were included into the model. Overall, 28 nominal model parameters were set non-zero specifying the presence of gene-TF interactions. Only one of those edges (ESAM  $\rightarrow$  TFAP4) received a negative weight and hence describes a repressing effect. The final model thus consists of 28 inferred gene-TF interactions and 102 TF-gene interactions, and 72 of the 121 genes are connected to at least one TF (figure 2, supplemental Cytoscape session file).

Network inference by TILAR outperformed all other algorithms tested in the benchmarking analysis described in more detail in the supplemental document. TILAR allowed for a higher

prediction accuracy than using just gene expression data or TFBS information alone, and performed best when incorporating text-mining information as well (adaptive TILAR). Therefore, we confirmed that the integrative modeling strategy is able to reconstruct GRNs more reliably than other established methods.

## Network characteristics

The inferred network is fairly complex but still readily interpretable due to the intuitive linear modeling scheme. Each network node is under control of only few regulators rendering the network sparse. The maximum in-degree in the GRN is 5 (on average 1.57). Nevertheless, some nodes (in particular TFs) are highly connected in the network, e.g. the NF- $\kappa$ B complex with an out-degree of 15 (figure 3B). Overall, most of the nodes are lowly connected and only a few are relatively highly connected. The node degrees are roughly distributed according to a power-law (scale-free topology), with the fraction  $P(k)$  of nodes in the network having  $k$  connections being estimated to follow  $P(k) \sim k^{-2.16}$ .

A closer look at the interactions in the model revealed gene sets co-regulated by a common TF (figure 3A). In the network, 9 genes are under transcriptional control of TFCP2, of which all but one were up-regulated after first week of IFN- $\beta$ -1a i.m. administration (e.g. ALOX12, GPR97 and CMTM5). Similarly, TF node ZIC1|ZIC2|ZIC3, which represents a group of ZIC family zinc finger proteins, is linked to 7 genes. Of these, 6 genes showed lowered mRNA levels in the patients' PBMC one week post therapy initiation in comparison to baseline (e.g. JUP, FAM26F and EGR2). The TFs ZIC1, ZIC2 and ZIC3 were grouped during TFBS analysis as they bind the same or at least very similar DNA sequences [38].

The GRN model contains 4 regulatory feedback loops and 25 feedforward loops of minimal length. An exemplary feedback loop in the reconstructed network is the regulatory chain "NF- $\kappa$ B  $\rightarrow$  EHD3  $\rightarrow$  TFCP2  $\rightarrow$  ALOX12  $\rightarrow$  NF- $\kappa$ B" (supplemental Cytoscape session file). EHD3 and ALOX12 were both found up-regulated immediately before second IFN- $\beta$  injection.

Patients with and without adverse reactions within the first 3 months of IFN- $\beta$  therapy showed significant differences in the transcript levels of the IFN- $\beta$ -responsive genes. There were 6 genes that not only displayed elevated amounts of mRNA after one-week therapy with IFN- $\beta$ -1a i.m., but that were also higher expressed before start of treatment in the group of patients with side effects (figure 4, supplemental table). Interestingly, 5 of these genes (ESAM, GNAZ, SLC24A3, ANKRD9, ALOX12) and 2 TFs (NF- $\kappa$ B, TOPORS) form a regulatory sub-network (figure 3B).

DNA-binding sites for NF-κB were found in the regulatory region of 3 of these genes. Our GRN inference result thus suggests that NF-κB regulates genes associated with adverse effects of IFN- $\beta$  therapy.

## Discussion

We analyzed transcriptional changes in response to 30 µg once-weekly, intramuscular IFN-β-1a treatment. A set of 121 genes was found significantly altered in the MS patients' PBMC during the first four weeks of drug administration. In this set, GO analysis revealed an overrepresentation of immunologically relevant genes (table 3). The impact of IFN-β therapy on gene expression was greater after one week of therapy than after one month, possibly suggesting a homeostatic response to the drug. We further determined the putative TFs mediating the gene regulatory effects of IFN-β and constructed a GRN model on the basis of the data. Subsequent network analysis identified "network motifs", i.e. frequent interaction patterns such as feedback and feedforward loops. A sub-network of genes was found more active in patients reporting side effects, demonstrating that the inferred molecular network could explain clinical heterogeneity.

Studies on pharmacogenomic effects of IFN-β in PBMC often differ in their experimental design. Gene expression changes have been investigated in response to different IFN-β treatment regimes (IFN-β-1a i.m. and s.c., as well as IFN-β-1b s.c.), sometimes pooling patients receiving any of these therapies. We examined a homogeneous cohort of patients who all were treated with IFN-β-1a i.m. at standard dose so that the findings are not distorted by differences in form, dose and route of drug administration. On the other hand, the observation period may range from a few hours until years into therapy. The dynamics during the first hours have been studied particularly intensively. It was shown that changes in expression can be seen within 2-4 h after drug administration [8-11]. Among the early IFN-β-induced genes are MX1, B2M, TNFSF10 and OAS1. Their mRNA and protein levels decline to baseline between 96 and 144 h after IFN-β administration [9-11,39,40]. Hence, once-a-week dosing of IFN-β-1a i.m. does not maintain these genes above baseline throughout a week. In our study, blood samples were drawn from the patients immediately before first, second and fifth intramuscular injection of IFN-β-1a. The data revealed the genes that are affected by the therapy even one week after last injection - presumably indirectly as a consequence of the activity of early IFN-induced genes - and therefore unveil the more persistent changes to the transcriptome.

Some of the 121 filtered genes have been mentioned in related studies on other IFN-β treatment regimes. For instance, OAS1, JUP and TNFSF10 have been described as up-regulated in PBMC of MS patients receiving IFN-β-1b s.c. for at least 6 months [41]. Apart from that, we observed an elevated expression of ESAM in agreement with the results of Annibali et al. [42]. In comparison to our study they used microarrays to obtain long-term expression profiles of PBMC from 7 MS

patients receiving IFN- $\beta$ -1a subcutaneously. We recently published a pharmacogenomic study where we applied Affymetrix microarrays to measure the PBMC expression levels of 25 patients treated with IFN- $\beta$ -1b s.c. [43]. In this data set, FCER1A was the only gene consistently down-regulated over the whole observation period of 2 years. FCER1A was also found significantly decreased after start of IFN- $\beta$ -1a i.m. therapy in the present work. Its down-regulation was confirmed by real-time PCR experiments in both studies. We thus conclude that FCER1A is generally repressed by IFN- $\beta$  treatment.

It is assumed that IFN- $\beta$  contributes to reduced lesion formation by modulating the inflammatory events at the blood-brain barrier. Our study provides some evidence that IFN- $\beta$  therapy improves the integrity of the blood-brain barrier via the activation of adhesion molecules and fibroblasts. Changes in the composition of adhesion molecules in the peripheral blood of MS patients receiving IFN- $\beta$  have already been discussed in the literature [44,45]. We found the two integrins ITGA2B and ITGB3 significantly up-regulated at one week into therapy. They are known to form a complex (the GPIb/IIIa complex), which mediates platelet aggregation by acting as a membrane receptor for fibrinogen and thus plays a crucial role in coagulation [46]. Up-regulated cell adhesion molecule ESAM that is under control of NF- $\kappa$ B in the GRN, as well as the down-regulated aminoacyl-tRNA synthetase WARS have been linked to the regulation of angiogenesis [47,48]. We further found an increased expression of FGFBP2, which might be important for fibroplasia and granulation tissue formation [49]. However, it remains an open question how these individual effects exactly contribute to the strengthening of the blood-brain barrier.

Evidently, the broad effects of IFN- $\beta$  are not purely anti-inflammatory and beneficial. IFN- $\beta$  treatment in relapsing-remitting MS can frequently induce systemic side effects such as flu-like symptoms with fever. A number of adverse effects emerge at the early phase after therapy initiation and then lessen over time, but a considerable amount of patients suffer from persistent or recurring side effects which may cause them to cease the treatment. To improve the clinical success of therapy an individualized tolerability management may be supported by molecular markers, but studies in this field are rare. Montalban et al. supposed that patients who develop fever during IFN- $\beta$  therapy generally have increased levels of IL-6 [50]. In our study, some (late) IFN- $\beta$ -responsive genes were found significantly higher expressed at baseline in the group of MS patients with side effects during the first 3 months of medication. Interestingly, NF- $\kappa$ B occurs as a regulator of a subset of these genes in the reconstructed GRN and its activity therefore potentially correlates with treatment-related adverse effects (figure 3B). Apart from that, HNF1A is connected to 7 genes in

the network. This TF is required for the expression of several liver-specific genes, and we hypothesize that its involvement coincides with liver function abnormalities common in patients who are treated with IFN- $\beta$  [51]. However, the clinical relevance of the differentially expressed genes and the potential roles of NF- $\kappa$ B and HNF1A remain to be further examined.

To better understand the pharmacologic effects of IFN- $\beta$  in MS, it is crucial to unravel the regulatory interaction structure of genes responsive to the immunotherapy. According to the GRN inferred by TILAR, IFN- $\beta$  mediates immune regulation through diverse TFs including some so far unrecognized in MS research. The inferred network reflects many fundamental properties that constitute a GRN including sparseness, scale-freeness, decentralization, self-regulation (feedback) and co-regulation. Moreover, the model reveals network regions, for instance genes down-regulated by ZIC family members and a highly interconnected NF- $\kappa$ B (figure 3), and thus provides testable hypotheses about the drug's mechanisms of action.

To date, our knowledge on the biological processes underlying MS and the efficacy and safety of available drugs is still limited. Individualized treatment and monitoring strategies are necessary to use these drugs in a more cost-effective way. An improved pharmacodynamic understanding at the transcript level should help to assess and individually optimize MS treatments with IFN- $\beta$ . Once established, this would protect the patients against unnecessary drug exposure and side effects. Several IFN- $\beta$ -responsive genes with potential prognostic value have already been discussed in the literature [35,41,52]. In this study, we described sustained PBMC gene expression changes in response to IFN- $\beta$ -1a i.m. and derived the regulating TFs. We identified a network region of genes associated to therapeutic side effects and linked to NF- $\kappa$ B activity. A potential regulatory feedback loop with NF- $\kappa$ B has been found. FCER1A, which is known to induce NF- $\kappa$ B [31], was affirmed as one of few genes significantly repressed by IFN- $\beta$ . We further discussed a set of genes that might reflect blood-brain barrier changes. To conclude, we showed that network analysis integrating different types of biological data could provide new insights into treatment-affected processes, and expose clinically relevant molecular differences in the patients. Supplementary information is available at The Pharmacogenomics Journal website.

## Acknowledgments

We deeply thank Peter Lorenz for helpful discussions and our lab assistants Gabriele Gillwaldt, Silvia Dilk, Ina Schröder and Ildikó Tóth for their help in performing the experiments. We are also grateful to study nurse Christa Tiffert for her invaluable contribution.

## **Conflicts of interest**

This study was partially funded by Biogen Idec. Prof Dr Zettl has received research support as well as speaking fees from Bayer, Biogen Idec, Merck Serono, Sanofi Aventis and Teva. Mr Hecker, Dr Goertsches, Mr Fatum, Dr Koczan, Prof Dr Thiesen and Dr Guthke declare no potential conflict of interest.

## References

- [1] Weiner HL. The challenge of multiple sclerosis: how do we cure a chronic heterogeneous disease? *Ann Neurol* 2009; **65**(3): 239-248.
- [2] Bradl M, Lassmann H. Progressive multiple sclerosis. *Semin Immunopathol* 2009; e-pub ahead of print 3 September 2009; doi:10.1007/s00281-009-0182-3.
- [3] Sospedra M, Martin R. Immunology of multiple sclerosis. *Annu Rev Immunol* 2005; **23**: 683-747.
- [4] Jacobs LD, Cookfair DL, Rudick RA, Herndon RM, Richert JR, Salazar AM et al. Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. *Ann Neurol* 1996; **39**(3): 285-294.
- [5] Jacobs LD, Beck RW, Simon JH, Kinkel RP, Brownscheidle CM, Murray TJ et al. Intramuscular interferon beta-1a therapy initiated during a first demyelinating event in multiple sclerosis. *N Engl J Med* 2000; **343**(13): 898-904.
- [6] Limmroth V, Malessa R, Zettl UK, Koehler J, Japp G, Haller P et al. Quality Assessment in Multiple Sclerosis Therapy (QUASIMS): a comparison of interferon beta therapies for relapsing-remitting multiple sclerosis. *J Neurol* 2007; **254**(1): 67-77.
- [7] Stark GR. How cells respond to interferons revisited: from early history to current complexity. *Cytokine Growth Factor Rev* 2007; **18**(5-6): 419-423.
- [8] Gilli F, Marnetto F, Caldano M, Sala A, Malucchi S, Di Sazio A et al. Biological responsiveness to first injections of interferon-beta in patients with multiple sclerosis. *J Neuroimmunol* 2005; **158**(1-2): 195-203.
- [9] Santos R, Weinstock-Guttman B, Tamaño-Blanco M, Badgett D, Zivadinov R, Justinger T et al. Dynamics of interferon-beta modulated mRNA biomarkers in multiple sclerosis patients with anti-interferon-beta neutralizing antibodies. *J Neuroimmunol* 2006; **176**(1-2): 125-133.
- [10] Weinstock-Guttman B, Bhasi K, Badgett D, Tamaño-Blanco M, Minhas M, Feichter J et al. Genomic effects of once-weekly, intramuscular interferon-beta1a treatment after the first dose and on chronic dosing: Relationships to 5-year clinical outcomes in multiple sclerosis patients. *J Neuroimmunol* 2008; **205**(1-2): 113-125.
- [11] Weinstock-Guttman B, Badgett D, Patrick K, Hartrich L, Santos R, Hall D et al. Genomic effects of IFN-beta in multiple sclerosis patients. *J Immunol* 2003; **171**(5): 2694-2702.
- [12] Waubant E, Vukusic S, Gignoux L, Dubief FD, Achiti I, Blanc S et al. Clinical characteristics of responders to interferon therapy for relapsing MS. *Neurology* 2003; **61**(2): 184-189.
- [13] Kappos L, Clanet M, Sandberg-Wollheim M, Radue EW, Hartung HP, Hohlfeld R et al. Neutralizing antibodies and efficacy of interferon beta-1a: a 4-year controlled study. *Neurology* 2005; **65**(1): 40-47.

- [14] Goertsches RH, Hecker M, Zettl UK. Monitoring of multiple sclerosis immunotherapy: from single candidates to biomarker networks. *J Neurol* 2008; **255 Suppl 6**: 48-57.
- [15] Comabella M, Martin R. Genomics in multiple sclerosis - current state and future directions. *J Neuroimmunol* 2007; **187**(1-2): 1-8.
- [16] Singh MK, Scott TF, LaFramboise WA, Hu FZ, Post JC, Ehrlich GD. Gene expression changes in peripheral blood mononuclear cells from multiple sclerosis patients undergoing beta-interferon therapy. *J Neurol Sci* 2007; **258**(1-2): 52-59.
- [17] Fernald GH, Knott S, Pachner A, Caillier SJ, Narayan K, Oksenberg JR et al. Genome-wide network analysis reveals the global properties of IFN-beta immediate transcriptional effects in humans. *J Immunol* 2007; **178**(8): 5076-5085.
- [18] Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models - a review. *Biosystems* 2009; **96**(1): 86-103.
- [19] Cho KH, Choo SM, Jung SH, Kim JR, Choi HS, Kim J. Reverse engineering of gene regulatory networks. *IET Syst Biol* 2007; **1**(3): 149-163.
- [20] McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann Neurol* 2001; **50**(1): 121-127.
- [21] Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M et al. Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics* 2007; **8**:446.
- [22] Hecker M, Goertsches RH, Engelmann R, Thiesen HJ, Guthke R. Integrative modeling of transcriptional regulation in response to antirheumatic therapy. *BMC Bioinformatics* 2009; **10**:262.
- [23] Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol* 2002; **3**(11): research0062.
- [24] Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics* 2007; **23**(2): 257-258.
- [25] Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 2007; **35**(Web Server issue): W253-W258.
- [26] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Statist* 2004; **32**(2): 407-499.
- [27] Killcoyne S, Carter GW, Smith J, Boyle J. Cytoscape: a community-based framework for network modeling. *Methods Mol Biol* 2009; **563**: 219-239.
- [28] Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. *SIAM Review* 2009; **51**(4): 661-703.

- [29] Melchjorsen J, Kristiansen H, Christiansen R, Rintahaka J, Matikainen S, Paludan SR et al. Differential Regulation of the OASL and OAS1 Genes in Response to Viral Infections. *J Interferon Cytokine Res* 2009; **29**(4): 199-208.
- [30] Ishibashi K, Kuwahara M, Gu Y, Tanaka Y, Marumo F, Sasaki S. Cloning and functional expression of a new aquaporin (AQP9) abundantly expressed in the peripheral leukocytes permeable to water and urea, but not to glycerol. *Biochem Biophys Res Commun* 1998; **244**(1): 268-274.
- [31] Kraft S, Kinet JP. New developments in FcepsilonRI regulation, function and inhibition. *Nat Rev Immunol* 2007; **7**(5): 365-378.
- [32] Giorelli M, Livrea P, Defazio G, Iacobelli L, Capobianco L, Picascia A et al. Interferon beta-1a counteracts effects of activation on the expression of G-protein-coupled receptor kinases 2 and 3, beta-arrestin-1, and regulators of G-protein signalling 2 and 16 in human mononuclear leukocytes. *Cell Signal* 2002; **14**(8): 673-678.
- [33] Hamilton JA. CSF-1 signal transduction. *J Leukoc Biol* 1997; **62**(2): 145-155.
- [34] Huang W-X, Huang MP, Gomes MA, Hillert J. Apoptosis mediators fasL and TRAIL are upregulated in peripheral blood mononuclear cells in MS. *Neurology* 2000; **55**(7): 928-934.
- [35] Wandinger KP, Lünemann JD, Wengert O, Bellmann-Strobl J, Aktas O, Weber A et al. TNF-related apoptosis inducing ligand (TRAIL) as a potential response marker for interferon-beta treatment in multiple sclerosis. *Lancet* 2003; **361**(9374): 2036-2043.
- [36] Kimberley FC, Screaton GR. Following a TRAIL: update on a ligand and its five receptors. *Cell Res* 2004; **14**(5): 359-372.
- [37] Tanaka N, Kawakami T, Taniguchi T. Recognition DNA sequences of interferon regulatory factor 1 (IRF-1) and IRF-2, regulators of cell growth and the interferon system. *Mol Cell Biol* 1993; **13**(8): 4531-4538.
- [38] Mizugishi K, Aruga J, Nakata K, Mikoshiba K. Molecular properties of Zic proteins as transcriptional regulators and their relationship to GLI proteins. *J Biol Chem* 2001; **276**(3): 2180-2188.
- [39] Stürzebecher S, Maibauer R, Heuner A, Beckmann K, Aufdembrinke B. Pharmacodynamic comparison of single doses of IFN-beta1a and IFN-beta1b in healthy volunteers. *J Interferon Cytokine Res* 1999; **19**(11): 1257-1264.
- [40] Buchwalder PA, Buclin T, Trinchard I, Munafò A, Biollaz J. Pharmacokinetics and pharmacodynamics of IFN-beta 1a in healthy volunteers. *J Interferon Cytokine Res* 2000; **20**(10): 857-866.
- [41] Stürzebecher S, Wandinger KP, Rosenwald A, Sathyamoorthy M, Tzou A, Mattar P et al. Expression profiling identifies responder and non-responder phenotypes to interferon-beta in multiple sclerosis. *Brain* 2003; **126**(Pt 6): 1419-1429.

- [42] Annibali V, Di Giovanni S, Cannoni S, Giugni E, Bompuzzi R, Mattei C et al. Gene expression profiles reveal homeostatic dynamics during interferon-beta therapy in multiple sclerosis. *Autoimmunity* 2007; **40**(1): 16-22.
- [43] Goertsches RH, Hecker M, Koczan D, Serrano-Fernandez P, Moeller S, Thiesen HJ et al. Long-term genome wide blood RNA expression profiles yield novel molecular response candidates for interferon beta-1b treatment in relapsing remitting multiple sclerosis. *Pharmacogenomics* (in press).
- [44] Kraus J, Bauer R, Chatzimanolis N, Engelhardt B, Tofighi J, Bregenzer T et al. Interferon-beta 1b leads to a short-term increase of soluble but long-term stabilisation of cell surface bound adhesion molecules in multiple sclerosis. *J Neurol* 2004; **251**(4): 464-472.
- [45] Floris S, Ruuls SR, Wierinckx A, van der Pol SM, Döpp E, van der Meide PH et al. Interferon-beta directly influences monocyte infiltration into the central nervous system. *J Neuroimmunol* 2002; **127**(1-2): 69-79.
- [46] Calvete JJ. On the structure and function of platelet integrin alpha IIb beta 3, the fibrinogen receptor. *Proc Soc Exp Biol Med* 1995; **208**(4): 346-360.
- [47] Ishida T, Kundu RK, Yang E, Hirata K, Ho YD, Quertermous T. Targeted disruption of endothelial cell-selective adhesion molecule inhibits angiogenic processes in vitro and in vivo. *J Biol Chem* 2003; **278**(36): 34598-34604.
- [48] Wakasugi K, Slike BM, Hood J, Otani A, Ewalt KL, Friedlander M et al. A human aminoacyl-tRNA synthetase as a regulator of angiogenesis. *Proc Natl Acad Sci U S A* 2002; **99**(1): 173-177.
- [49] Ogawa K, Tanaka K, Ishii A, Nakamura Y, Kondo S, Sugamura K et al. A novel serum protein that is selectively produced by cytotoxic lymphocytes. *J Immunol* 2001; **166**(10): 6404-6412.
- [50] Montalban X, Durán I, Río J, Sáez-Torres I, Tintoré M, Martínez-Cáceres EM. Can we predict flu-like symptoms in patients with multiple sclerosis treated with interferon-beta? *J Neurol* 2000; **247**(4): 259-262.
- [51] Tremlett H, Oger J. Hepatic injury, liver monitoring and the beta-interferons for multiple sclerosis. *J Neurol* 2004; **251**(11): 1297-1303.
- [52] Comabella M, Lünemann JD, Río J, Sánchez A, López C, Julià E et al. A type I interferon signature in monocytes is associated with poor response to interferon-beta in multiple sclerosis. *Brain* 2009; **132**(Pt 12): 3353-3365.

## Tables

**Table 1** : Blood expression profiling studies on intramuscular IFN- $\beta$ -1a treatment in MS.

Author	Year	#Patients	Sample	Microarray	GEO Accession	Sampling	#Genes ↑↓
This study	2010	24	PBMC	Affymetrix HG-U133 A and B	GSE19285	0h, 1W, 1M	121
Weinstock-Guttman et al. [10,11]	2008, 2003	22	PBL	GeneFilters GF211 DNA arrays	not available	0h, 8x within 1W, 1M, 6M, 12M	1539
Singh et al. [16]	2007	5	PBMC	CodeLink UniSet Human I Bioarray	GSE5574	0h, 24h, ~6M	136
Fernald et al. [17]	2007	2	Whole blood	UCSF Human 21k Oligo array	GSE5678	0h, up to 7x within 1W	~1000

Four microarray studies have been conducted in this field since 2003. The table provides the number of patients examined and the sample material used to measure mRNA levels. Moreover, the blood sampling time-points as well as the number of genes found modulated in expression during therapy (on the basis of different filtering criteria, "#Genes ↑↓") are shown. GEO = gene expression omnibus, PBL = peripheral blood lymphocytes, h = hour, W = week, M = month.

**Table 2** : Clinical and demographic characteristics of the patients.

Patient ID	Gender	Age (years)	EDSS (baseline)	EDSS (3 months)	Relapse during first 3 months	Side effects during first 3 months
Pat01	Female	47	2.5	3.0	No	No
Pat02	Female	45	1.5	1.5	No	Yes
Pat03	Female	30	2.5	2.5	No	Yes
Pat04	Female	40	1.0	1.0	No	Yes
Pat05	Female	28	1.5	1.5	No	No
Pat06	Female	31	1.0	1.0	No	Yes
Pat07	Female	44	1.0	1.0	No	Yes
Pat08	Female	45	1.5	1.5	No	Yes
Pat09	Female	24	0.0	1.0	No	Yes
Pat10	Male	44	1.5	1.5	No	Yes
Pat11	Male	43	1.0	1.0	No	Yes
Pat12	Male	27	1.0	1.0	No	No
Pat13	Female	39	0.0	0.0	No	No
Pat14	Male	22	0.0	0.0	No	Yes
Pat15	Female	34	0.0	0.0	No	Yes
Pat16	Female	20	1.5	1.5	Yes	Yes
Pat17	Female	30	1.0	1.0	No	Yes
Pat18	Female	38	0.5	1.0	No	Yes
Pat19	Male	41	1.5	1.5	Yes	Yes
Pat20	Female	35	0.0	0.0	No	Yes
Pat21	Female	38	2.0	1.5	No	Yes
Pat22	Female	49	0.0	0.0	No	Yes
Pat23	Male	26	1.0	1.0	No	Yes
Pat24	Female	40	1.5	1.5	No	Yes
<b>Mean</b>		35.8	1.0	1.1		
<b>SD</b>		8.3	0.7	0.7		

SD = standard deviation.

**Table 3** : Overrepresented terms of the GO biological process ontology.

GO BP ID	GO Term	P-value	Expected Count	Count
GO:0006955	immune response	<0.0001	3.89	14
GO:0002376	immune system process	0.0001	5.34	16
GO:0009605	defense response	0.0005	4.47	13
GO:0030334	I-kappaB kinase/NF-kappaB cascade	0.0019	0.53	4
GO:0009611	positive regulation of immune system process	0.0025	2.92	9
GO:0050896	positive regulation of immune response	0.0026	15.03	26
GO:0051270	protein kinase cascade	0.0028	0.59	4
GO:0007218	inflammatory response	0.0028	0.59	4
GO:0051046	regulation of immune system process	0.0030	0.60	4
GO:0051047	regulation of immune response	0.0034	0.30	3
GO:0042108	positive regulation of multicellular organismal process	0.0036	0.31	3
GO:0006954	fatty acid metabolic process	0.0046	2.07	7
GO:0006916	immunoglobulin mediated immune response	0.0062	1.16	5
GO:0042221	B cell mediated immunity	0.0078	4.11	10
GO:0051050	activation of immune response	0.0086	0.42	3
GO:0009967	lymphocyte mediated immunity	0.0090	1.28	5

*P*-values were computed for each GO term based on the hypergeometric distribution. Only functional categories with *P*-value<0.01 and where at least 3 out of the 121 IFN- $\beta$ -responsive genes are associated ("Count") are shown. "Expected count" gives the expected number of genes in the list of filtered genes to be found at each tested category term. BP = biological process.

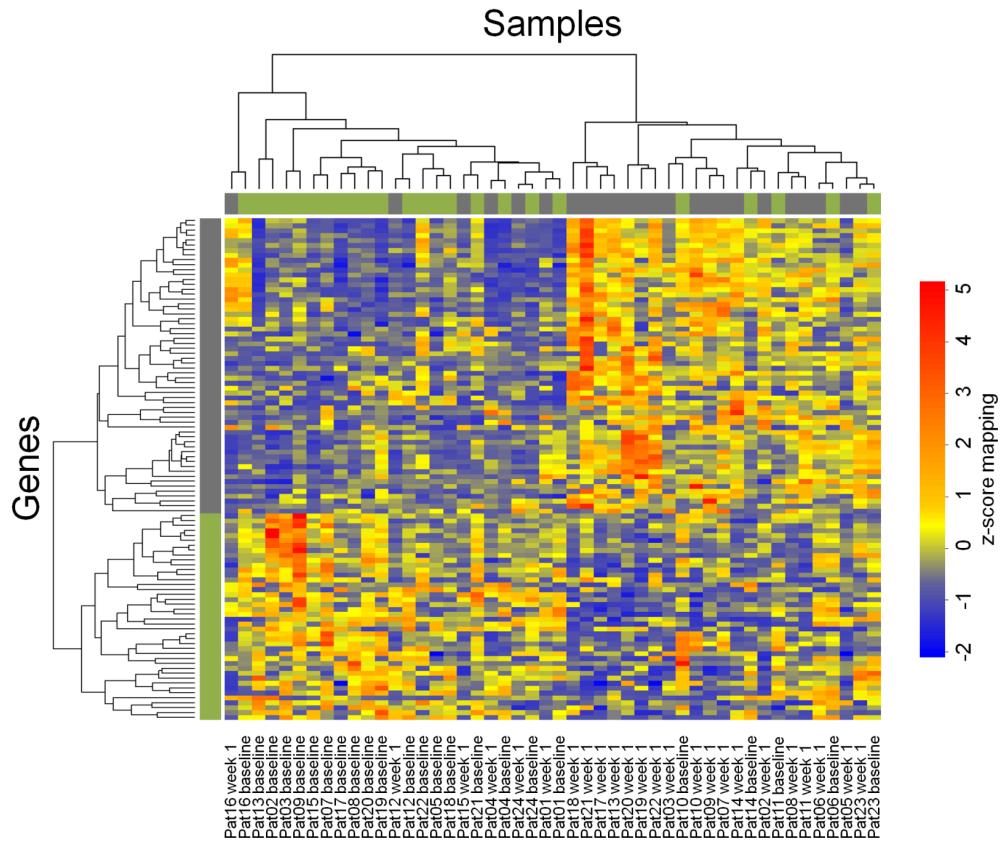
**Table 4** : TFBS overrepresented in the regulatory regions of up- and down-regulated genes.

TF Symbol	Transfac Accession	Official Full Name	P-value	Expected Count	Count
SRF	M00152, M00186, M00215	serum response factor (c-fos serum response element-binding transcription factor)	0.003	8.22	17
TFCP2	M00072	transcription factor CP2	0.019	5.37	11
TOPORS	M00480	topoisomerase I binding, arginine-serine-rich	0.019	5.40	11
HOXA9, MEIS1, TGIF1	M00418, M00419, M00420, M00421	homeobox A9, Meis homeobox 1, TGFB-induced factor homeobox 1	0.022	12.46	20
HNF1A	M00132, M00206	HNF1 homeobox A	0.040	3.85	8
TBP	M00216, M00252, M00471	TATA box binding protein	0.081	6.84	11
IRF1, IRF2	M00062, M00063	interferon regulatory factor 1 and 2	0.086	5.33	9
ZIC1, ZIC2, ZIC3	M00448, M00450, M00449	Zic family members 1-3	0.091	9.48	14
MEF2A	M00006, M00232, M00231, M00233, M00026	myocyte enhancer factor 2A	0.094	7.86	12
TFAP4	M00175, M00176, M00005	transcription factor AP-4 (activating enhancer binding protein 4)	0.097	13.01	18
NFKB1, NFKB2, REL, RELA	M00051, M00052, M00053, M00054, M00194, M00208	nuclear factor of kappa light polypeptide gene enhancer in B-cells family members	0.098	15.68	21
<b><math>\Sigma = 152</math></b>					

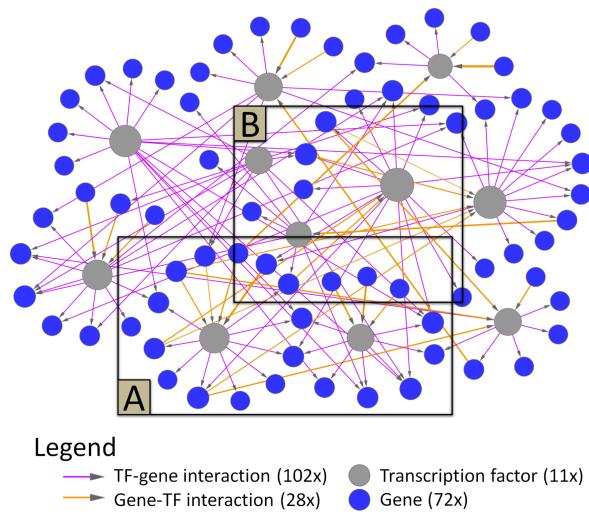
Evolutionarily conserved TFBS of 11 TF entities were found enriched. The list includes IRFs and NF-κB, which are known TFs in IFN-β signaling [7]. Highly similar Transfac motifs were consolidated before calculating the *P*-values. The column "Count" shows the number of genes that possess a DNA-binding site for the respective TF. In sum, there are 152 predicted TF-gene interactions.

## Figure legends

**Figure 1** : Clustering analysis and graphical representation of the data.

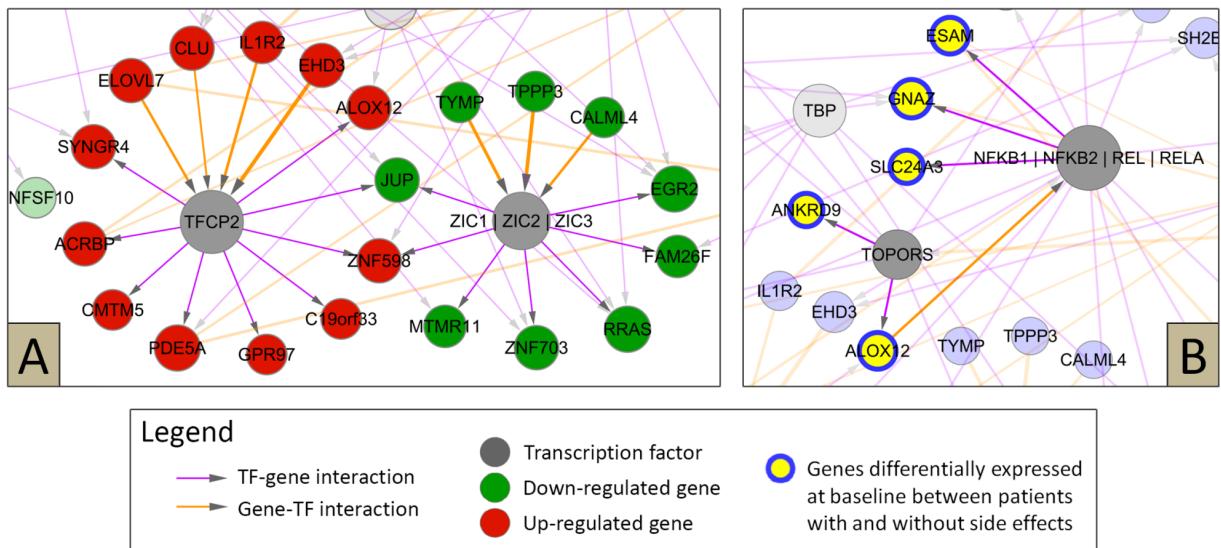


Heatmap of 102 genes identified as up- (60 genes, gray labeled rows) or down-regulated (42 genes, green) one week after first intramuscular injection of IFN- $\beta$ -1a. Hierarchical clustering was performed based on the complete linkage method and Pearson's correlation coefficient as a measure of similarity. Signal intensities were centered and scaled row-wise (yielding z-scores) for visualization purposes. Despite a strong interindividual variability, the clustering tends to separate baseline measurements (green labeled columns) from gene expression levels obtained at one week into therapy (gray). The row labels of the heatmap are given in the supplemental table.

**Figure 2 :** Gene regulatory network inferred by TILAR.

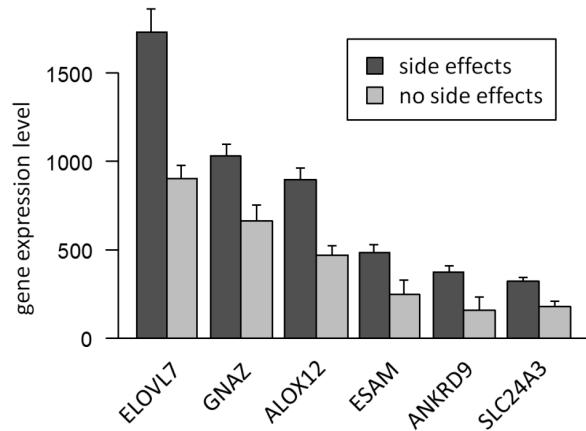
Gene expression data and TFBS predictions were utilized to build a GRN model of 102 TF-gene and 28 gene-TF interactions. The network comprises 72 of the 121 genes that were found up- or down-regulated during therapy. The remaining genes are not shown here. Two parts of the network are presented in detail in figure 3. The full model is available as a Cytoscape session file in the supplementary material.

**Figure 3** : Detail views of the GRN model shown in figure 2.



**(A)** Network inference by TILAR takes into account that TF target genes are often co-expressed. All but one of the genes that are regulated by TFCP2 are up-regulated after therapy initiation. Similarly, a set of down-regulated genes is predicted to contain DNA-binding sites for ZIC family members in their regulatory regions. Both TFs receive multiple regulatory inputs. **(B)** Some of the filtered genes have potential relevance to treatment-related adverse effects. The figure on the right shows a sub-network of genes that are higher expressed at baseline in patients experiencing side effects. Three of these genes are regulated by NF- $\kappa$ B according to the model. Outer parts are shown with lower opacity.

**Figure 4 :** Expression differences between patients with and without side effects.



Of the 121 filtered genes, 6 genes were found significantly higher expressed ( $t$ -test  $P$ -value $<0.05$ ) before treatment in patients suffering from side effects early in therapy. Bars indicate mean and standard error.

## Supplemental data legends

**Supplemental table** (*supplemental\_table.xls*, XLS Excel spreadsheet): List of 121 genes up- or down-regulated during first four weeks of intramuscular IFN- $\beta$ -1a treatment. Genes with transcriptional changes in response to therapy initiation were identified by use of the MAID filtering method and a test of statistical significance. There were 72 genes higher expressed and 49 genes lower expressed at week 1 or 4 versus baseline in the PBMC of the 24 MS patients in our study. The table provides diverse types of information for each gene, e.g. Entrez ID, official full name and the calculated MAID-scores.

**Supplemental document** (*supplemental\_document.pdf*, PDF document): This document is divided into two parts. The first part is on the validation of the microarray data by real-time PCR measurements. In the second part, we describe how we evaluated the inference quality of the TILAR algorithm using text-mining information.

**Supplemental Cytoscape session file** (*supplemental\_cytoscape\_file.cys*, CYS Cytoscape session file): Cytoscape session file of the inferred GRN model. The network model describes the regulatory interactions between TFs and the genes with expression changes during first month of IFN- $\beta$  administration. A simplified visualization of the network is shown in figure 2, while detail views are shown in figure 3.

This file could not be submitted due to the Journal's submission restrictions. Therefore, we provide it at our institute's web page (provisional URL):

<http://www.hki-jena.de/index.php/0/2/495/download/2822>

## 7. Discussion

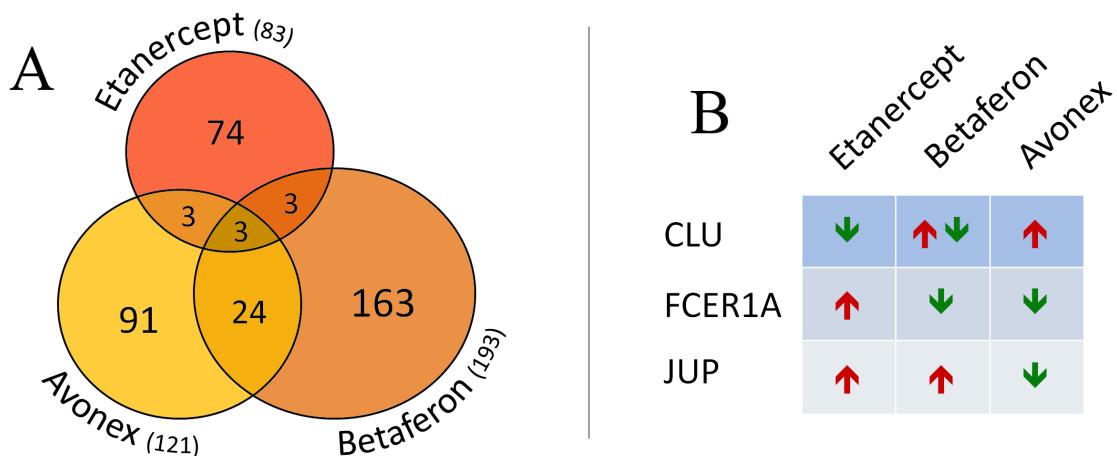
In the works presented here, transcriptional effects were studied in response to three different autoimmune disease therapies. The starting point of each analysis was a genome-wide RNA expression dataset. These measurements provided expression levels in blood for RA and MS patients at different therapy time points: immediately before first drug injection (baseline), as well as after a few days, weeks or even after two years of immunomodulatory treatment. The (regulatory) interactions between genes up- or down-regulated during each particular therapy were examined by use of integrative GRN modeling and network analysis.

The implementation of methods for Affymetrix microarray analysis and the development of the integrative GRN inference algorithm TILAR were main parts of this work. In the following, the biological results from manuscript II-IV are briefly reviewed and compared among the different therapies. Afterwards, the methods are discussed in more detail and an outlook is given on network inference and its use in computational medicine.

### 7.1. Discussion of main results

When comparing to baseline transcript levels, stronger gene expression changes were always observed relatively early in therapy, suggesting adaptations that occur as a homeostatic response to the drugs. Functional analysis of the up- and down-regulated genes revealed an overrepresentation of immunologically relevant genes: For all three therapies, genes associated to the Gene Ontology (GO) terms "immune system process", "immune response" and "defense response" were significantly enriched in the sets of filtered genes. However, genes annotated with these high-level GO terms may accomplish quite diverse immune functions, and only few genes were found influenced by each of the three investigated biologic agents. As shown in figure 3A, three genes were significantly modulated not only in response to Etanercept in RA patients, but also in response to both IFN- $\beta$  preparations in MS patients: junction plakoglobin (JUP), clusterin (CLU), and Fc IgE receptor, alpha polypeptide (FCER1A). All three genes were differentially regulated between the therapies (figure 3B) and it is therefore difficult to judge their general therapeutic role. Nevertheless, common cellular and molecular components are presumably influenced by the treatments. As could be expected, there is a considerable overlap (27 genes) when comparing the lists of filtered genes of the both MS therapies.

In the pharmacodynamic studies on Etanercept and Avonex, the regulatory regions of the genes with significant expression changes were screened for overrepresented evolutionarily



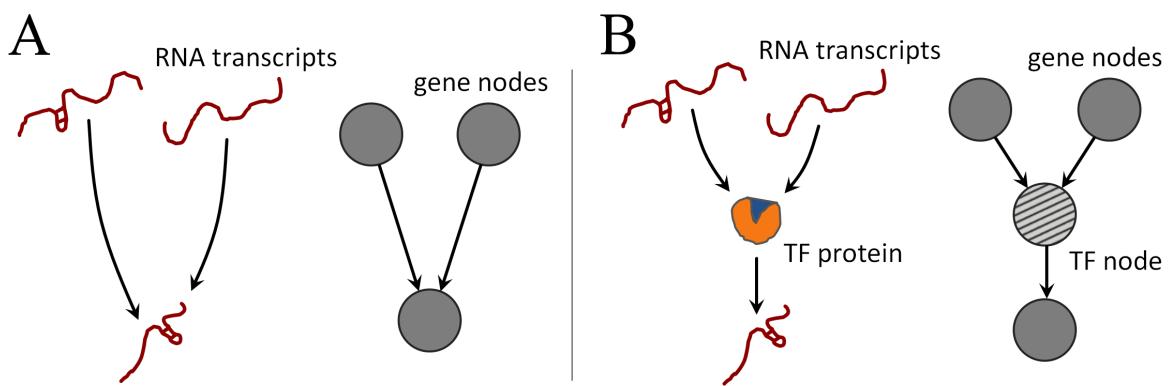
**Figure 3.** Several genes were found modulated in more than one of the three studies. **(A)** Overall, 83 genes were up- or down-regulated during first week of Etanercept therapy in RA patients. Similarly, 193 and 121 genes were filtered by analyzing the expression profiles of MS patients receiving Betaferon and Avonex, respectively. For the latter two, the first month into treatment was considered here for better comparability, even if in the Betaferon study gene expression was also measured at later time points (after one and two years). Three genes occur in each of the gene lists. **(B)** These three genes are differentially regulated in response to the different treatments (up-regulation: red arrow, down-regulation: green arrow). However, the therapeutic effects can only be compared qualitatively rather than quantitatively. The studies are about different therapies to different diseases, and as their comparison was not an issue from the very beginning, the obtained microarray data were analyzed independently, e.g. using different chip definitions and gene filtering criteria.

conserved TFBS. In this way, TFs were identified that putatively play a role in mediating the drugs' gene regulatory effects. For instance, the DNA sequence motif for CCAAT/enhancer binding protein beta (CEPB), a TF that has been associated to chronic inflammation in RA [27], was found for 11 out of the 83 genes responsive to Etanercept. Binding sites for NF- $\kappa$ B and the IFN regulatory factors 1 and 2, which are known TFs in IFN- $\beta$  signaling [26], were overrepresented in the promoter region of genes modulated during first month of IFN- $\beta$ -1a therapy. TATA-like elements and binding sites for Zic family members (ZIC) 1-3 as well as hepatic nuclear factor 1 homeobox A (HNF1A) were found enriched in both analyses. This again indicates that the therapies might share some immunoregulatory activities.

The information on overrepresented TFBS was then used to reconstruct the regulatory interactions between respective TFs and the filtered genes using the novel TILAR inference method. TILAR stands for TFBS-integrating least angle regression. "TFBS-integrating" here means that the method integrates gene expression data and TF binding predictions to infer a linear GRN model, and the "least angle regression" (commonly abbreviated LARS) is an

efficient regression algorithm that was employed to determine the model parameters on the basis of the data [28]. TILAR considers genes and TFs as two distinct types of nodes in the network and defines directed TF-gene and gene-TF interactions as network edges (figure 4). Only the transcript levels of the genes are required for learning the model. TF protein activities are described implicitly by the genes that influence the TFs. TILAR constrains the model to include a TF-gene interaction only when the TF is predicted to bind the regulatory region of the gene and is thus presumably involved in its transcriptional regulation. Hence, TF-gene interactions can be interpreted as true physical interactions between TF proteins and gene promoter DNA sequences. In contrast, gene-TF interactions can have very different meanings and may represent indirect effects (e.g. through a signaling cascade). They are therefore difficult to augment by external information in general. For reverse-engineering the GRNs relevant to Etanercept and Avonex therapy no prior knowledge on gene-TF interactions was incorporated. However, it has been demonstrated in both studies that if such knowledge would be available to a certain degree (e.g. from text-mining), it can be utilized to further increase the quality of the inferred networks (adaptive TILAR). Moreover, in both works, TILAR and its adaptive variant outperformed all other tested inference algorithms that utilize either gene expression data or TFBS information alone. The performance was evaluated using gene interaction information provided by the software PathwayArchitect 2.0.1. PathwayArchitect contains molecular relationships which have been extracted from biomedical literature using text-mining and different curated biological databases. Data-driven modeling by TILAR was not (yet) applied to the Betaferon dataset. Here, filtered gene lists were just imported into PathwayArchitect to retrieve interactions between the genes modulated during long-term therapy.

As a result, relatively large gene interaction networks were built in each study. The networks reveal insights into gene regulatory processes that may play a crucial role in the treatment of RA and relapsing-remitting MS, and thus are useful for generating new and testable hypotheses on the drugs' molecular mechanisms of action. To further scrutinize the transcriptional effects in particular early in therapy, respective complex interaction patterns were then analyzed in more detail. Sparseness / scale-freeness, decentralization, self-regulation and co-regulation are characteristic for the inferred networks. These features are discussed to enhance a GRN's robustness in terms of structural stability [4,29]. Some genes that are annotated to the same GO term share a common TFBS in the models. An example was given in manuscript II, where four network genes belonging to "immune system process" are all regulated in a TATA-dependent manner. Beyond, to some extent, the GRNs reconstructed by TILAR could show clinical heterogeneity among the patients: Genes



**Figure 4.** The GRN model from figure 1 and its representation according to the TILAR modeling approach. **(A)** Figure 1B revisited: When inferring a network from gene expression data only, one seeks for influences between RNA transcripts. In this case, edges in the model not necessarily correspond to true/causal molecular relationships. **(B)** In comparison, TILAR incorporates TFBS predictions and constrains genes to regulate other genes via TFs. Interactions between genes are then implicitly defined in the network by gene-TF and TF-gene interactions. In this exemplary model, two genes control the activity of a TF and therefore the transcription of its target gene, which is a bit closer to the more realistic regulation scheme shown in figure 1A.

differentially expressed between responders and non-responders were connected to the same TF (NF-1) in the Etanercept RA therapy study. Similarly, in the Avonex study, an NF- $\kappa$ B-centered network region of IFN- $\beta$ -responsive genes was found to coincide with therapeutic side effects. This demonstrated that improving the knowledge on gene regulation during therapy and correlating expression differences in the network with clinical data may disclose signatures that could allow the prognosis of beneficial and adverse responses to the drugs.

## 7.2. Discussion of methods

### 7.2.1. Experimental approach

In each of the presented works, gene expression profiles were analyzed for a homogeneous cohort of patients who all were prescribed the same treatment (co-medications ignored) at the same clinic. The microarray data used in manuscript II provided expression levels for 19 RA patients. Each patient was given a standard dose of 25 mg Etanercept subcutaneously twice weekly, and blood was taken before first, second and third drug injection, i.e. at baseline as well as after 3 and 6 days. The pharmacodynamic effects of the both IFN- $\beta$  preparations were investigated separately to account for differences in dose and route of drug administration, and their different immunogenic properties [30]. For instance, the frequency of injections differs: Avonex is given once a week, while Betaferon is given every other day.

Blood samples were again drawn always immediately before injection, i.e. two days and one week after previous IFN- $\beta$  injection in case of Betaferon and Avonex, respectively. In comparison to Betaferon, which is given subcutaneously, the serum concentrations of the latter (IFN- $\beta$ -1a) may be sustained after its intramuscular administration due to prolonged absorption from the injection site. Though, as discussed in manuscript IV, the list of genes with expression changes during Avonex therapy is expected to contain many genes indirectly modulated as a consequence of the activity of early IFN- $\beta$ -responsive genes. This also explains the certain but relatively small overlap (27 genes) when comparing the lists of filtered genes of both IFN- $\beta$  treatments (figure 3A).

Another important consideration is that the microarray data revealed the gene expression levels of PBMC. The PBMC fraction contains monocytes and lymphocytes, the latter comprise natural killer cells (NK cells), T cells and B cells. Therefore, even if PBMC generally allow to immunologically assess therapeutic effects, one should be aware that the contribution of the different cell types was neglected in the studies. In consequence, the reconstructed molecular networks represent complex intracellular as well as intercellular interactions. Ongoing research efforts aim to further fractionate the immune cell populations to analyze the behavior of specific cells. Besides, to better understand autoimmune disease processes and the mechanisms of immunomodulatory drugs, it is also crucial to investigate the role of the target tissues, in connection with the infiltrated cells of the immune system. As the presented analyses followed an immune-centered view and focused on the expression alterations of immune cells, the mutual interactions between immune system and target tissue factors (e.g. of the synovium in case of RA) were out of scope.

### 7.2.2. Microarray data preprocessing

Before reconstructing the regulatory interactions between genes up- or down-regulated during therapy via TILAR, the microarray data were preprocessed using custom chip definition files (CDF). The original Affymetrix chip definitions contain probes which do not match any transcript and probes which cross-hybridize to transcripts of multiple genes. They also represent many genes by more than one probeset, which often leads to discordant expression signals for the same transcript. Custom CDF provide an improved annotation of Affymetrix probesets and realize a one-to-one correspondence to genes [31]. In consequence, when the data are preprocessed by use of a custom CDF, genes and their expression levels can be unequivocally assigned to nodes in the network model. In general, this makes it easier to extend and validate the interaction network with external molecular

information from biological databases. In addition, it not only alleviates an integrative GRN modeling, but also eases the interpretation of such networks. In the pharmacogenomic studies on anti-TNF- $\alpha$  and intramuscular IFN- $\beta$  treatment, up-to-date custom CDF that are based on information contained in the GeneAnnot database were used to process the raw probe intensities [31,32]. Custom CDF for microarrays for other organisms than human are available elsewhere [33].

Next, genes differentially expressed during therapy compared to baseline were selected to infer the underlying regulatory interactions from the data. As described in more detail in manuscript I, there is a strong relationship between network complexity (i.e. network size and level of detail of the model), the amount of data required for inference and the quality of the results. In the presented works, the self-developed MAID filtering approach in combination with a paired *t*-test statistic was used to filter genes with significant expression changes during treatment. MAID considers the shape of MAS5.0-processed (and loess normalized) Affymetrix microarray data and yields signal intensity-dependent fold-changes (see also section "Methods" in manuscript II). Afterwards, the TILAR algorithm was applied to reconstruct the network of 83 filtered genes measured on 55 microarrays in the Etanercept study, and the network of 121 genes (72 microarrays) in the study on the effects of Avonex administration. Since the TILAR strategy generally reduces the number of free linear model parameters (see discussion below), the respective amount of expression data was considered sufficient for modeling such medium scale GRNs. From the mathematical perspective, the inference of even larger networks is possible, but would certainly generate structures that are less reliable and more difficult to study further.

### 7.2.3. Integrative network inference

Network inference by TILAR is based on the linear modeling approach. In comparison to other modeling formalisms, linear models have a number of benefits: they allow to describe the direction of interactions as well as feedback and feedforward loops, and take into account that gene regulators act in combination. Moreover, as regulatory relationships are specified quantitatively, no prior discretization of the data is necessary (as e.g. for Boolean networks). A disadvantage is that linear models are only crude linearizations of the true system as in fact gene regulation is a dynamic non-linear process including saturation and stochasticity. However, as a system of non-linear equations consists of much more model parameters, and because the "true" form of these equations is usually not known, linear models are often preferred for modeling GRNs. In general, linear functions are appropriate to capture the

main features of a GRN when similar conditions have been measured in the microarray experiments [34].

The major advantage of linear models is that computationally very efficient (and deterministic) algorithms exist to estimate the model parameters, i.e. to fit the model to the given data. TILAR utilizes the Lasso (least absolute shrinkage and selection operator) method for linear regression [35]. The Lasso minimizes the residual sum of squares subject to the sum of the absolute value of the model parameters being less than a constant (bound  $s$ ). In dependence on this tuning value  $s$ , this constraint produces some parameters that are exactly zero and hence gives sparse models. The implementation of TILAR employs the efficient LARS procedure for fitting the entire Lasso sequence (i.e. for all values of  $s$ ) with the cost of a single ordinary least squares (OLS) fit [28]. It is also possible to compute all Lasso solutions if the system of linear equations is underdetermined, that is, if there are more model parameters ( $p$ ) than equations ( $n$ ). As a drawback, there is a conflict of optimal parameter setting (low fitting error) and consistent parameter selection in the Lasso, which led to the use of a LARS/OLS hybrid in TILAR [28]. Other regression concepts have been proposed to better solve this conflict, e.g. the relaxed Lasso and the elastic net [36,37]. The elastic net has the capacity of selecting groups of highly correlated  $p$  and is particularly suited for the  $p \gg n$  case. The strength of the elastic net was observed in a DREAM 2008 (dialogue on reverse-engineering assessment and methods) challenge, where the best performing algorithm in predicting missing gene expression measurements utilized the elastic net [38]. However, relaxed Lasso and elastic net define in their constraint function a second tuning value in addition to  $s$ , which needs to be determined e.g. by cross-validation and therefore increases the computational efforts.

TILAR is an integrative GRN modeling approach that merges information from different types of data to increase the chance of inferring true regulatory effects. It combines the RNA expression data with TFBS predictions and, if available, literature-mining information (adaptive TILAR). A soft integration is the concept to include the additional but also uncertain information [34]. That means prior knowledge on TF-gene and gene-TF edges is used to increase the probability for an edge, and not merely as a filter. TILAR incorporates only a subset of the predicted TF-gene interactions obtained from TFBS analysis. This was realized by a backward stepwise selection procedure that iteratively applies LARS for respective subsets until a local error minimum is found. As a remark, although a single LARS calculation takes only a few seconds, this backward selection can be computationally demanding if the network to be inferred is very large. Besides, in the adaptive TILAR,

literature evidence from PathwayArchitect is softly integrated by putting less penalty on model parameters that represent gene-TF interactions suggested by text-mining than on those not suggested. Due to the TFBS-integrating constraint in TILAR, all target genes of a particular TF receive the same regulatory input. More importantly, the number of free model parameters is generally reduced:  $p$  is bound by the product of the number of genes and the number of TFs in the GRN. It is therefore feasible to reconstruct relatively large networks.

The inference quality of TILAR not only depends on the quality of the gene expression data, but also on the accuracy of TFBS (and text-mining) information. For human, experimental evidence for TFBS is still scarce and scattered, but different databases provide computationally predicted TFBS. Most of them rely on known binding motifs searched in the genome and filtered for cross-species sequence conservation and/or motif combinations (i.e. co-localized TFBS clusters). Examples for such databases are UCSC tfbsConsSites (that was used in the studies), cisRED, SwissRegulon and oPOSSUM [39-42]. The latter three also contain TFBS for other organisms than human, but limit the predictions to predefined promoter regions, and all differ in their contents. Moreover, not all predicted TFBS refer to biologically functional binding sites as TF-DNA bindings depend on several factors, e.g. nucleosome occupancy. A further issue is that often more than one (similar) binding motif is associated to a particular TF in these databases. However, sufficient prior knowledge on gene-TF interactions is even more difficult to obtain because of their variable meanings. In the proposed adaptive TILAR, text-mining information was integrated as prior knowledge. A text-mining approach is able to detect indirect gene interactions, but a drawback is that such information (independently of the software used) is always incomplete and biased (there are *per se* less interactions described in the literature for less prominent genes), error-prone (in particular concerning the direction of interactions), and not disease- or cell type-specific.

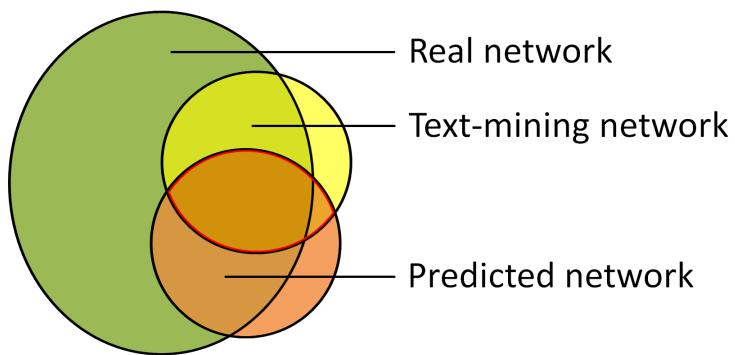
#### 7.2.4. Evaluation of inference performance

The performance evaluation of a GRN inference method, i.e. the assessment of sensitivity and specificity of reconstructed interactions, is a fundamental problem in the field. Only first attempts have been made to establish standardized benchmarking systems and criteria. Notably, the DREAM initiative is dedicated to the understanding of limitations and strengths of methods for inferring networks from biological data [43]. The project aim is to create gold standards for network inference by providing datasets and defining common objective evaluation measures. Since 2007, DREAM arranges annually competitions on reverse-

engineering networks and on predicting expression levels using networks. The true structure of these networks (and/or additional experimental data) is known to the organizers, but hidden to the community. Research groups are asked to submit the results of their algorithms, thereby getting the opportunity to assess and compare them with the performance of others.

However, TILAR has not been evaluated within DREAM since so far the challenges did not (or only in parts) include the issue of integration of various types of data. For real network data gene names were sometimes withheld and *in silico* data can not be augmented by additional information at all. The performance of TILAR has therefore been rated by comparing the inferred networks (i.e. the respective path of models produced by the algorithm, from an empty to a fully connected network) with gene-gene interactions extracted from PathwayArchitect (manuscript II and the supplemental material to manuscript IV given in the appendix). For this analysis, genes well-described in literature were selected so that PathwayArchitect delivered hundreds of interactions. As TILAR distinguishes genes and TFs, the information integrated during modeling was different from the information used to evaluate the method. Performance measures that were also proposed in DREAM [44], namely RPC (recall-precision curves) and ROC (receiver operating characteristic) curves, have been applied in the benchmarking. As a result, TILAR achieved superior accuracy compared to other GRN inference algorithms in both benchmarking studies, which independently demonstrated the benefit of an integrative modeling. Still, for a general two-class prediction problem, the attained areas under the RPC and ROC curves might seem rather modest. However, this is somewhat misleading as discussed in more detail in manuscript I. A main reason for "low" performance lies in the text-mining information used for the assessment. This is incomplete and contains errors, and thus is only a "bronze standard". In consequence, even a GRN inference method which delivers almost the "true" network would not reach areas under these curves close to 1 (figure 5).

Still, the outcome of a GRN reconstruction should always be taken with some care and validated by arguments outside the analysis. This may include the confirmation of expression changes (e.g. for selected genes), the verification of regulatory effects postulated in the network (e.g. the DNA-binding of certain TFs), the determination of protein level alterations (e.g. for TFs in the GRN) and new cell type-specific *in vitro* perturbation assays to quantify time-dependent responses. The GRN models thus may be a starting point for additional experiments.



**Figure 5.** Relationships among inferred, text-mining-based, and real network. The text-mining network (in each of both independent benchmarking analyses) was composed of PathwayArchitect-retrieved regulatory gene interactions and contains some incorrectly assigned false positives. This network is a subset of the actual (genome-wide and multi-level) transcriptional network, represented in the figure as the "real network". The network reconstructed using TILAR or another algorithm ("predicted network") also includes falsely identified connections. A small overlap of the predicted and the text-mining network leads to relatively low performance measures.

## 7.3. Open issues and outlook

### 7.3.1. Prediction of clinical responses

The three drugs, for which the blood transcriptional responses were investigated, are well established in the therapy of RA and MS today. Nevertheless, the frequency of injections, the side effects that particularly occur at the beginning of therapy, and the treatment failures are all reasons to search for optimized treatment strategies. The ultimate goal is to identify, as early as possible, the patients that will benefit the most from a specific medication. This would eventually allow for a more individualized treatment and disease monitoring (personalized medicine), and increase both clinical efficacy and cost-effectiveness of the therapies. The identification of patients with a favorable prognosis must also consider potential side effects. A better drug tolerability usually leads to enhanced compliance, and thus can positively affect long-term clinical outcomes [45,46].

Different therapeutic responses might be reflected in transcriptional differences between the patients, possibly in the magnitude of transcriptional changes early in therapy. Therefore, an improved pharmacodynamic understanding can help to individually optimize autoimmune disease therapies. Molecular markers to discriminate "good" from "poor" responders and to discern patients at high risk for (persistent) adverse effects have already been proposed in the literature for RA [47,48] and MS [49-51]. In perspective, transcript analyses of such genes

with prognostic value could become part of a routine laboratory screening and serve as input for algorithm-based risk-benefit prediction. In the studies on Etanercept and Avonex (manuscript II and IV), coherent sub-networks were found, which are enriched of genes differentially expressed between patient subgroups (according to clinical response and side effects, respectively). However, the true clinical relevance of these genes remains to be further examined.

### 7.3.2. Further development of TILAR

One can think of many extensions and possible improvements to the integrative GRN modeling using TILAR. For instance, the modeling framework can be adapted to infer dynamic models given appropriate time-course data (i.e. with short intervals between two successive measurements). Such a dynamic model can be interpreted as a linearization of the true non-linear regulatory functions around a working point (first-order approximation). As outlined in the introduction, the model then corresponds to a system of linear difference equations and assumes equidistant intervals between the time points. If non-equidistant data are given, one might therefore interpolate intermediary time points before inference (see also section "data requirements" in manuscript I). Dynamic linear models allow to simulate and analyze the behavior of the network. An eigenvalue analysis of matrix  $W$  (that contains the gene-gene edge weights) can show whether the network is stable in time against random fluctuations (dynamical stability). Though, so far, no efficient algorithm for learning this kind of models exists that includes dynamical robustness as a modeling constraint.

A linear model may be dynamically unstable just because stabilizing non-linear effects are not captured. Other researchers therefore introduced non-linear functions, while still utilizing the efficient mathematical techniques for fitting linear models [52,53]. For instance, a sigmoid influence function may be used for dynamic models to take into account saturation effects. Such non-linearity can be simply realized by transforming the gene expression data accordingly. The trick is that the reverse-engineering problem becomes linear on the transformed data. If the model is then used for simulation studies, the model outputs must be retransformed again. However, the use of (possibly coupled) non-linear terms might be in conflict with experimental data requirements and can increase the risk of model overfitting and complicate model interpretation.

Furthermore, the TILAR method can be modified to softly integrate knowledge on (binary) protein-protein interactions (PPI), e.g. from the IntAct database [54]. Here, one could follow an approach that was applied by Nariai et al. [55] and Bonneau et al. [56]: In brief, the idea

is to describe PPI in the network model by use of virtual nodes corresponding to protein complexes. This can be done by inserting additional columns (model parameters) to the regression matrix of the regression problem that is solved by LARS. The final model then might (softly) include a protein complex if this allows for a better fit to the expression data. PPI were not incorporated in the presented studies, because only one protein hetero-dimer was found in the list of anti-TNF- $\alpha$ -responsive genes (EBI-360918: NFKBIA-TUBA1A) and none for the genes transcriptionally modulated by IFN- $\beta$ .

Other biological databases can be used either to obtain prior evidence on gene-TF interactions or as an alternative way to evaluate the inference performance. For instance, one can use publicly available microarray data, e.g. from the Gene Expression Omnibus database [57]. In general, the quality of external gene expression data is difficult to assess. Besides, the data comprise various experimental conditions that are not related to RA and MS disease conditions and where different regulatory processes are active. Although the benefit may thus be limited, it is possible to softly integrate such data during GRN inference, in TILAR by defining lower penalties for gene-TF edges if the external data support a transcriptional correlation between the gene and the TF or the gene and a gene putatively controlled by the TF. Microarray databases also contain TF binding data (e.g. from tiling arrays) which can provide useful information on TFBS (and therefore TF-gene edges). The capability of including public microarray data into the learning of linear GRN models via a Lasso-type shrinkage was shown by Gustafsson et al. [38].

A more difficult challenge is to integrate signaling pathway information, which would mean to link signaling networks with GRNs – from a historical point of view two distinct classes of cellular networks. The major difficulties here are that many different pathway databases with different data formats exist [58] and that the contained signaling cascades are rarely associated to the transcription of specific gene sets. Two efforts should be mentioned that promise to ease the central access and the use of such information in the future. First, the Pathway Commons project which aims to provide a convenient point of access to biological pathway information collected from public pathway databases [59]. Second, ConsensusPathDB, a (meta-)resource of curated human signaling pathways that also lists genes transcriptionally regulated by the pathways [60]. When using such databases, respective gene relationships (even indirect ones, i.e. if two genes are connected via an interaction path) may be integrated by TILAR as prior knowledge on gene-TF or TF-gene interactions.

Further extensions in TILAR are conceivable: 1) A scale-free constraint might be more powerful than the sparseness constraint realized by the Lasso. However, no efficient algorithm to consider scale-freeness during reconstruction of linear models has been developed yet. 2) The backward stepwise selection step in TILAR is a bit time-consuming and might be improved by use of other heuristics or by prioritizing predicted TF-gene interactions by additional criteria (e.g. motif score or distance to the transcription start site). 3) Penalties for known gene-TF edges might be set lower the more often this interaction can be found in the literature-mined information or if an molecular biology expert considers this interaction as particularly important.

## 7.4. Concluding remarks

Manuscript II-IV demonstrate the use of network analysis to provide novel insights into functional mechanisms of immunomodulatory therapies for RA and MS. The newly developed TILAR approach, which allows to integrate different types of data during GRN modeling, may find applicability in a wider context, e.g. to infer gene regulatory interactions relevant in other conditions or even in other organisms. The reconstructed networks yield useful information on the genes responsive to the respective treatments, and the TFs that are assumed to mediate their transcriptional regulation. The networks show the existence of hubs and network motifs (e.g. feedback loops), and a scale-free distribution of node degrees. They also describe sub-networks of genes co-regulated by a common TF and genes with functional similarity (e.g. genes known to participate in certain immune system processes). In addition, specific network regions expose expression differences of possible clinical relevance in the patients. Therefore, this work is one further step towards understanding the molecular mechanisms underlying the effects of biologic agents in RA and MS patients.

Still, there is a gap to close between molecular events and clinical outcomes. Little is known about the precise factors that account for beneficial and adverse therapeutic effects in the individual patient. To increase the efficacy of available drugs it is important to identify and establish predictive biomarkers that allow for a more individualized medication and treatment monitoring. In future, such markers may guide clinicians in choosing the optimal, tailored treatment dependent on the patient's molecular and clinical attributes. Besides, novel drugs will offer improved mechanisms of targeting autoimmune diseases while minimizing side effects. As a single drug seems less likely to entirely block a polygenic complex disease like RA or MS, combination therapies may be the key for a more successful treatment.

These promising perspectives require an improved knowledge on the biological processes related to these diseases. Integrative bioinformatics in the field of computational medicine, which here includes the investigation of GRNs, can help to better understand the biology of human diseases and their treatment. In the long term, advancements in bioinformatics and biotechnology will therefore allow to achieve improvements in therapeutic efficacy and safety.

## 8. Summary

Autoimmune diseases are disorders where an aberrant immune response leads the body to attack its own cells and organs. Two common diseases with an autoimmune basis are rheumatoid arthritis (RA) and multiple sclerosis (MS). Their complex genetic background in concert with environmental factors causes considerable differences between the patients in clinical presentation and response to treatments. The introduction of immunomodulatory biologic drugs significantly improved the treatment of these diseases. For instance, TNF- $\alpha$ -antagonists (e.g. Etanercept) and IFN- $\beta$  (e.g. Betaferon and Avonex) have been proven to reduce symptoms and progression of RA and relapsing-remitting MS, respectively. However, to date, our knowledge on the underlying molecular mechanisms and the efficacy and safety of these drugs is still limited.

In this work, the transcriptional effects induced by these three therapies were studied to provide a deeper molecular understanding of the drugs' immunotherapeutic mechanisms. The analyses included network inference techniques to reconstruct the gene regulatory networks (GRNs) of therapy-responsive genes. The starting point of each analysis was an Affymetrix DNA microarray dataset. For groups of patients, the data provided gene expression levels in peripheral blood mononuclear cells (PBMC) immediately before first and selected subsequent drug administrations. This allowed to investigate the general changes in expression within the first days (Etanercept), weeks (Avonex) and years (Betaferon) of therapy. Newly devised MAID scores were utilized to filter genes that characterize the biological response to the treatments. After start of each therapy, many genes were significantly up- or down-regulated compared to their expression before therapy initiation. Most of them are known to participate in immune system processes. The transcription factors (TFs) that are putatively involved in the regulation of these genes were ascertained by screening the genes' regulatory regions for overrepresented TF binding sites (TFBS) (Etanercept and Avonex study only). Several TFBS were detected, e.g. as could be expected, DNA-binding sites for IFN regulatory factors (IRF) were enriched for IFN- $\beta$ -1a-modulated genes.

Afterwards, interactions between the genes with significant expression changes during therapy were examined. In the study on the effects of Betaferon treatment, gene interaction networks were built using literature-mining information. In the other two studies, a self-developed network inference algorithm called TILAR was used to deduce a model of the underlying GRN from the gene expression data. TILAR distinguishes genes and TFs in the network and mathematically describes the mutual regulatory effects between them by a

system of linear equations. TILAR is an integrative modeling strategy that constrains the model to only include a TF-gene interaction, when the TFBS overrepresentation analysis predicts that the TF binds at the regulatory region of the gene. The method is also able to (softly) incorporate evidence on gene-TF interactions, if some genes are known to control TF activities (adaptive TILAR). According to the presented approach, genes are supposed to regulate other genes in a linear additive relationship indirectly via one or more TFs. This results in a relatively low number of free model parameters and makes the models straightforward to interpret. Two benchmarking analyses demonstrated that TILAR reconstructs GRNs more reliably than other established inference algorithms. Moreover, the integrative modeling allows for a higher prediction accuracy than using just gene expression data or TFBS information alone. Apart from that, TILAR is adaptable and extensible and thus may be applied also in other applications.

As a result, complex interaction structures were obtained, which describe the gene regulatory effects in response to each of the three treatments and thus provide useful hypotheses about the drugs' molecular mechanisms of action. The networks show a self- and co-regulatory organization and a scale-free topology. Moreover, the GRNs inferred by TILAR were analyzed in the context of clinical data: An NF- $\kappa$ B-linked gene sub-network was found higher expressed in MS patients with side effects to intramuscular IFN- $\beta$ -1a administration. Similarly, a network region was lower expressed in clinical responders in the pharmacodynamic study on Etanercept RA therapy.

These findings showed that GRN analysis integrating different types of data as well as prior biological knowledge can contribute to the investigation of treatment-associated processes, and reveal different expression patterns in the patients. Implications of clinical relevance have been discussed for RA and MS therapy. Further studies are required to establish early indicators of long-term therapeutic outcomes and potential adverse effects. In future, new techniques in biotechnology and bioinformatics and the growing amount of available experimental data should allow for more comprehensive network models of transcriptional regulation, and thus help to study (individual) molecular therapeutic effects in more detail.

## 9. Zusammenfassung

Autoimmunerkrankungen sind Krankheiten, bei denen sich eine fehlgeleitete Immunantwort gegen körpereigene Zellen und Organe richtet. Unter anderem werden die Rheumatoide Arthritis (RA) und die Multiple Sklerose (MS) als Autoimmunkrankheiten eingestuft. Ein komplexes, bisher unverstandenes Zusammenspiel aus Umweltfaktoren und genetischer Veranlagung spielt bei der Pathogenese dieser beiden Krankheiten eine entscheidende Rolle und führt dazu, dass Krankheitsausprägung und -verlauf von Patient zu Patient recht verschieden sind. Zur Behandlung dieser Krankheiten werden auch immunmodulatorische biologische Medikamente eingesetzt. So können TNF- $\alpha$ -Blocker (z.B. Etanercept) im Falle der RA und IFN- $\beta$ -Präparate (z.B. Betaferon und Avonex) im Falle der schubförmig remittierenden MS Krankheitssymptome lindern und das Fortschreiten der Erkrankung verzögern. Bis heute sind jedoch die zugrunde liegenden molekularen Mechanismen dieser Therapien weitestgehend unbekannt. Zudem ist ihre Wirksamkeit und Verträglichkeit bei einigen Patienten unzureichend.

In dieser Arbeit wurde die transkriptionelle Antwort auf die drei genannten Therapien untersucht, um die Wirkungsweise dieser biologischen Medikamente besser zu verstehen. Dabei wurden Methoden der Netzwerkinferenz eingesetzt, mit dem Ziel, die genregulatorischen Netzwerke (GRNs) der in ihrer Expression veränderten Gene zu rekonstruieren. Ausgangspunkt dieser Analysen war jeweils ein Affymetrix-Mikroarray-Datensatz. Diese Daten lieferten für eine Gruppe von Patienten Informationen über die Genexpression peripherer mononukleärer Blutzellen (PBMC) zu verschiedenen Therapiezeitpunkten: unmittelbar vor der ersten Injektion sowie vor ausgewählten nachfolgenden Medikamentengaben. Der Beobachtungszeitraum erstreckte sich dabei über mehrere Tage (Etanercept), Wochen (Avonex) bzw. Jahre (Betaferon). Mit Hilfe der neu entwickelten MAID-Scores wurden die Gene gefiltert, die nach Therapiebeginn besonders stark hoch- bzw. herunterreguliert sind. Es stellte sich heraus, dass die Mehrheit dieser Gene bei Immunsystemprozessen beteiligt ist. In der Avonex- und Etanercept-Studie wurden anschließend die genregulatorischen Regionen dieser Gene auf überrepräsentierte Transkriptionsfaktor-Bindestellen (TFBS) analysiert, um so auf Transkriptionsfaktoren (TF) zu schließen, die potentiell deren Expression maßgeblich regulieren.

Anschließend wurde untersucht, welche wechselseitigen Interaktionen zwischen den Genen bestehen, die unter Therapie moduliert werden. In der Betaferon-Studie wurden dazu entsprechende Geninteraktionsnetzwerke mit einer Text-Mining-Software erstellt. In den

anderen beiden Studien wurde ein neu implementierter Netzwerkinferenz-Algorithmus (TILAR) verwendet, um GRN-Modelle direkt aus den Genexpressionsdaten abzuleiten. TILAR unterscheidet grundsätzlich zwischen Genen und TF und beschreibt die regulatorischen Effekte zwischen diesen durch ein lineares Gleichungssystem. TILAR verfolgt dabei eine integrative Modellierungsstrategie, da nur solche TF-Gen-Interaktionen erlaubt sind, für die die TFBS-Überrepräsentationsanalyse ergab, dass der TF in der regulatorischen Region des Gens binden kann. Die Methode erlaubt auch Vorwissen über bekannte Gen-TF-Interaktionen adaptiv einzubeziehen. Diese bilden ab, dass ein Gen (möglicherweise indirekt) die Aktivität eines TFs kontrolliert. Vorteile dieses linearen, integrativen Modellierungsansatzes sind zum einen die relativ geringe Anzahl freier Modellparameter und zum anderen die einfache Interpretierbarkeit der resultierenden Modelle. Zwei unabhängige Benchmark-Tests zeigten, dass TILAR eine bessere Netzwerkrekonstruktion ermöglicht als andere Inferenzverfahren. Die Modellgüte ist höher, als wenn nur Genexpressionsdaten oder nur TFBS-Vorhersagen für die Inferenz der genregulatorischen Interaktionen verwendet werden würden.

Im Ergebnis wurden komplexe Netzwerkstrukturen rekonstruiert, welche die regulatorischen Beziehungen zwischen den Genen beschreiben, die im Verlauf der Therapien differentiell exprimiert sind. Aus diesen lassen sich neue Hypothesen über die Wirkungsweise der Medikamente ableiten. Die Netzwerke beinhalten Rückkopplungsmechanismen und besitzen eine skaleninvariante Topologie. Die mit TILAR berechneten GRNs wurden weiterhin zusammen mit klinischen Daten ausgewertet. Hier zeigte sich in der Avonex-Studie, dass NF- $\kappa$ B möglicherweise ein Teilnetz reguliert, das Gene enthält, die bei MS-Patienten mit deutlichen Nebenwirkungen höher exprimiert sind. Analog wurde eine GRN-Region gefunden, die ein niedriges Expressionsniveau bei RA-Patienten aufweist, die sehr gut auf die Therapie mit Etanercept ansprechen.

Die Analyse von GRNs kann zu einem besseren Verständnis Therapie-relevanter Prozesse beitragen und transkriptionelle Unterschiede zwischen den Patienten aufzeigen. Die Modellberechnung kann dabei verschiedene Typen von Daten sowie biologisches Vorwissen integrativ berücksichtigen. Weitere gezielte Studien sind nötig, um Biomarker zu etablieren, die die langfristige klinische Wirkung und Verträglichkeit von RA- und MS-Therapien frühzeitig und individuell vorhersagen können. Neue biotechnologische und bioinformatische Techniken sowie die zunehmende Menge verfügbarer experimenteller Daten sollte in naher Zukunft umfassendere genregulatorische Netzwerkmodelle ermöglichen und somit helfen, (individuelle) molekulare therapeutische Effekte genauer zu erforschen.

## References

- [1] D'haeseleer P, Wen X, Fuhrman S, Somogyi R: **Linear modeling of mRNA expression levels during CNS development and injury.** *Pac Symp Biocomput.* 1999, **4**:41-52.
- [2] van Someren EP, Wessels LF, Backer E, Reinders MJ: **Genetic network modeling.** *Pharmacogenomics* 2002, **3**(4):507-525.
- [3] Gardner TS, Faith JJ: **Reverse-engineering transcription control networks.** *Physics of Life Reviews* 2005, **2**:65-88.
- [4] Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet.* 2004, **5**(2):101-113.
- [5] Benson M, Breitling R: **Network theory to understand microarray studies of complex diseases.** *Curr Mol Med.* 2006, **6**(6):695-701.
- [6] Hewagama A, Richardson B: **The genetics and epigenetics of autoimmune diseases.** *J Autoimmun.* 2009, **33**(1):3-11.
- [7] Jacobson DL, Gange SJ, Rose NR, Graham NM: **Epidemiology and estimated population burden of selected autoimmune diseases in the United States.** *Clin Immunol Immunopathol.* 1997, **84**(3):223-243.
- [8] Cush JJ, Weinblatt ME, Kavanaugh A: *Rheumatoid Arthritis: Early Diagnosis and Treatment.* 2nd edition. West Islip: Professional Communications; 2008.
- [9] Compston A, Coles A: **Multiple sclerosis.** *Lancet* 2008, **372**(9648):1502-1517.
- [10] Raine CS, McFarland HF, Hohlfeld R: *Multiple Sclerosis: A Comprehensive Text.* Philadelphia: Saunders Elsevier; 2008.
- [11] Edwards JC, Cambridge G, Abrahams VM: **Do self-perpetuating B lymphocytes drive human autoimmune disease?** *Immunology* 1999, **97**(2):188-196.
- [12] Bergsteinsdottir K, Yang HT, Pettersson U, Holmdahl R: **Evidence for common autoimmune disease genes controlling onset, severity, and chronicity based on experimental models for multiple sclerosis and rheumatoid arthritis.** *J Immunol.* 2000, **164**(3):1564-1568.
- [13] Somers EC, Thomas SL, Smeeth L, Hall AJ: **Are individuals with an autoimmune disease at higher risk of a second autoimmune disorder?** *Am J Epidemiol.* 2009, **169**(6):749-755.
- [14] Kieseier BC, Hartung HP: **Current disease-modifying therapies in multiple sclerosis.** *Semin Neurol.* 2003, **23**(2):133-146.

- [15] Smolen JS, Steiner G: **Therapeutic strategies for rheumatoid arthritis.** *Nat Rev Drug Discov.* 2003, **2**(6):473-488.
- [16] McInnes IB, Schett G: **Cytokines in the pathogenesis of rheumatoid arthritis.** *Nat Rev Immunol.* 2007, **7**(6):429-442.
- [17] Sospedra M, Martin R: **Immunology of multiple sclerosis.** *Annu Rev Immunol.* 2005, **23**:683-747.
- [18] Elliott MJ, Maini RN, Feldmann M, Kalden JR, Antoni C, Smolen JS, Leeb B, Breedveld FC, Macfarlane JD, Bijl JA, Woody JN: **Randomised double-blind comparison of chimeric monoclonal antibody to tumour necrosis factor alpha (cA2) versus placebo in rheumatoid arthritis.** *Lancet* 1994, **344**(8930):1105-1110.
- [19] Bathon JM, Martin RW, Fleischmann RM, Tesser JR, Schiff MH, Keystone EC, Genovese MC, Wasko MC, Moreland LW, Weaver AL, Markenson J, Finck BK: **A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis.** *N Engl J Med.* 2000, **343**(22):1586-1593.
- [20] The Lenercept Multiple Sclerosis Study Group and The University of British Columbia MS/MRI Analysis Group: **TNF neutralization in MS: results of a randomized, placebo-controlled multicenter study.** *Neurology* 1999, **53**(3):457-465.
- [21] Jacobs LD, Cookfair DL, Rudick RA, Herndon RM, Richert JR, Salazar AM, Fischer JS, Goodkin DE, Granger CV, Simon JH, Alam JJ, Bartoszak DM, Bourdette DN, Braiman J, Brownscheidle CM, Coats ME, Cohan SL, Dougherty DS, Kinkel RP, Mass MK, Munschauer FE 3rd, Priore RL, Pullicino PM, Scherokman BJ, Weinstock-Guttman B, Whitham RH, The Multiple Sclerosis Collaborative Research Group (MSCRG): **Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis.** *Ann Neurol.* 1996, **39**(3):285-294.
- [22] Limmroth V, Malessa R, Zettl UK, Koehler J, Japp G, Haller P, Elias W, Obhof W, Viehöver A, Meier U, Brosig A, Hasford J, Putzki N, Kalski G, Wernsdörfer C, QUASIMS Study Group: **Quality Assessment in Multiple Sclerosis Therapy (QUASIMS): a comparison of interferon beta therapies for relapsing-remitting multiple sclerosis.** *J Neurol.* 2007, **254**(1):67-77.
- [23] Kraus J, Oschmann P: **The impact of interferon-beta treatment on the blood-brain barrier.** *Drug Discov Today* 2006, **11**(15-16):755-762.
- [24] Waubant E, Vukusic S, Gignoux L, Dubief FD, Achiti I, Blanc S, Renoux C, Confavreux C: **Clinical characteristics of responders to interferon therapy for relapsing MS.** *Neurology* 2003, **61**(2):184-189.
- [25] Feldmann M, Maini RN: **Anti-TNF alpha therapy of rheumatoid arthritis: what have we learned?** *Annu Rev Immunol.* 2001, **19**:163-196.

- [26] Stark GR: **How cells respond to interferons revisited: from early history to current complexity.** *Cytokine Growth Factor Rev.* 2007, **18**(5-6):419-423.
- [27] Pope RM, Lovis R, Mungre S, Perlman H, Koch AE, Haines GK 3rd: **C/EBP beta in rheumatoid arthritis: correlation with inflammation, not disease specificity.** *Clin Immunol.* 1999, **91**(3):271-282.
- [28] Efron B, Hastie T, Johnstone I, Tibshirani R: **Least angle regression.** *Ann Statist* 2004, **32**(2):407-499.
- [29] Lesne A: **Robustness: confronting lessons from physics and biology.** *Biol Rev Camb Philos Soc.* 2008, **83**(4):509-532.
- [30] Gneiss C, Tripp P, Reichartseder F, Egg R, Ehling R, Lutterotti A, Khalil M, Kuenz B, Mayringer I, Reindl M, Berger T, Deisenhammer F: **Differing immunogenic potentials of interferon beta preparations in multiple sclerosis patients.** *Mult Scler.* 2006, **12**(6):731-737.
- [31] Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli GA, Bicciato S: **Novel definition files for human GeneChips based on GeneAnnot.** *BMC Bioinformatics* 2007, **8**:446.
- [32] Chalifa-Caspi V, Yanai I, Ophir R, Rosen N, Shmoish M, Benjamin-Rodrig H, Shkler M, Stein TI, Shmueli O, Safran M, Lancet D: **GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes.** *Bioinformatics* 2004, **20**(9):1457-1458.
- [33] Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res.* 2005, **33**(20):e175.
- [34] Gustafsson M, Hörnquist M: **Integrating various data sources for improved quality in reverse engineering of gene regulatory networks.** In *Handbook of Research on Computational Methodologies in Gene Regulatory Networks.* 1st edition. Edited by: Das S, Caragea D, Welch SM, Hsu WH. New York: Medical Information Science Reference; 2009:476-496.
- [35] Tibshirani R: **Regression shrinkage and selection via the lasso.** *J R Statist Soc B* 1996, **58**(1):267-288.
- [36] Meinshausen N: **Relaxed Lasso.** *Comput Statist Data Anal.* 2007, **52**(1):374-393.
- [37] Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Statist Soc B* 2005, **67**(2):301-320.

- [38] Gustafsson M, Hörnquist M: **Gene Expression Prediction by Soft Integration and the Elastic Net - Best Performance of the DREAM3 Gene Expression Challenge.** *PLoS One* 2010, **5**(2):e9134.
- [39] Weirauch M, Raney B: **TFBS conserved track at UCSC genome browser** [<http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=tfbsConsSites>]
- [40] Robertson AG, Bilenky M, Lin K, He A, Yuen W, Dagpinar M, Varhol R, Teague K, Griffith OL, Zhang X, Pan Y, Hassel M, Sleumer MC, Pan W, Pleasance ED, Chuang M, Hao H, Li YY, Robertson N, Fjell C, Li B, Montgomery SB, Astakhova T, Zhou J, Sander J, Siddiqui AS, Jones SJM: **cisRED: A database system for genome scale computational discovery of regulatory elements.** *Nucleic Acids Res.* 2006, **34**(Database issue):D68-D73.
- [41] Pachkov M, Erb I, Molina N, van Nimwegen E: **SwissRegulon: a database of genome-wide annotations of regulatory sites.** *Nucleic Acids Res.* 2007, **35**(Database issue):D127-D131.
- [42] Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW: **oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.** *Nucleic Acids Res.* 2005, **33**(10):3154-3164.
- [43] Stolovitzky G, Monroe D, Califano A: **Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference.** *Ann N Y Acad Sci.* 2007, **1115**:1-22.
- [44] Stolovitzky G, Prill RJ, Califano A: **Lessons from the DREAM2 Challenges.** *Ann N Y Acad Sci.* 2009, **1158**:159-195.
- [45] Garcia-Gonzalez A, Richardson M, Garcia Popa-Lisseanu M, Cox V, Kallen MA, Janssen N, Ng B, Marcus DM, Reveille JD, Suarez-Almazor ME: **Treatment adherence in patients with rheumatoid arthritis and systemic lupus erythematosus.** *Clin Rheumatol.* 2008, **27**(7):883-889.
- [46] Klauer T, Zettl UK: **Compliance, adherence, and the treatment of multiple sclerosis.** *J Neurol.* 2008, **255**(Suppl 6):87-92.
- [47] Koczan D, Drynda S, Hecker M, Drynda A, Guthke R, Kekow J, Thiesen HJ: **Molecular discrimination of responders and nonresponders to anti-TNF alpha therapy in rheumatoid arthritis by etanercept.** *Arthritis Res Ther.* 2008, **10**(3):R50.
- [48] Hughes LB, Danila MI, Bridges SL: **Recent advances in personalizing rheumatoid arthritis therapy and management.** *Personalized Medicine* 2009, **6**(2):159-170.
- [49] Montalban X, Durán I, Río J, Sáez-Torres I, Tintoré M, Martínez-Cáceres EM: **Can we predict flu-like symptoms in patients with multiple sclerosis treated with interferon-beta?** *J Neurol.* 2000, **247**(4):259-262.

- [50] Baranzini SE, Mousavi P, Rio J, Caillier SJ, Stillman A, Villoslada P, Wyatt MM, Comabella M, Greller LD, Somogyi R, Montalban X, Oksenberg JR: **Transcription-based prediction of response to IFNbeta using supervised computational methods.** *PLoS Biol.* 2005, **3**(1):e2.
- [51] Comabella M, Lünemann JD, Río J, Sánchez A, López C, Julià E, Fernández M, Nonell L, Camiña-Tato M, Deisenhammer F, Caballero E, Tortola MT, Prinz M, Montalban X, Martin R: **A type I interferon signature in monocytes is associated with poor response to interferon-beta in multiple sclerosis.** *Brain* 2009, **132**(Pt 12):3353-3365.
- [52] Weaver DC, Workman CT, Stormo GD: **Modeling regulatory networks with weight matrices.** *Pac Symp Biocomput.* 1999, 112-123.
- [53] Gustafsson M, Hörnquist M, Lundström J, Björkegren J, Tegnér J: **Reverse engineering of gene networks with LASSO and nonlinear basis functions.** *Ann NY Acad Sci.* 2009, **1158**:265-275.
- [54] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risso J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H: **IntAct - open source resource for molecular interaction data.** *Nucleic Acids Res.* 2007, **35**(Database issue):D561-D565.
- [55] Nariai N, Kim S, Imoto S, Miyano S: **Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks.** *Pac Symp Biocomput.* 2004, 336-347.
- [56] Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V: **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.** *Genome Biol.* 2006, **7**(5):R36.
- [57] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res.* 2009, **37**(Database issue):D885-D890.
- [58] Bader GD, Cary MP, Sander C: **Pathguide: a pathway resource list.** *Nucleic Acids Res.* 2006, **34**(Database issue):D504-D506.
- [59] Bader B, Cerami E, Demir E, Gross B, Sander C: **Pathway Commons** [<http://www.pathwaycommons.org>]
- [60] Kamburov A, Wierling C, Lehrach H, Herwig R: **ConsensusPathDB - a database for integrating human functional interaction networks.** *Nucleic Acids Res.* 2009, **37**(Database issue):D623-D628.

## Appendix

### **Network analysis of transcriptional regulation in response to intramuscular interferon-beta-1a multiple sclerosis treatment**

#### Supplementary Material

Michael Hecker<sup>1</sup>, Robert Hermann Goertsches<sup>2,3</sup>, Christian Fatum<sup>3</sup>, Dirk Koczan<sup>2</sup>, Hans-Juergen Thiesen<sup>2</sup>, Reinhard Guthke<sup>1</sup>, Uwe Klaus Zettl<sup>3</sup>

<sup>1</sup> Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knoell-Institute,  
Beutenbergstr. 11a, D-07745 Jena, Germany,  
e-mail: {michael.hecker, reinhard.guthke}@hki-jena.de,  
phone: +49-3641-5321083, fax: +49-3641-5320803

<sup>2</sup> University of Rostock, Institute of Immunology, Schillingallee 70, D-18055 Rostock, Germany,  
e-mail: {dirk.koczan, hans-juergen.thiesen}@med.uni-rostock.de

<sup>3</sup> University of Rostock, Department of Neurology, Gehlsheimer Str. 20, D-18147 Rostock,  
Germany, e-mail: {robert.goertsches, christian.fatum, uwe.zettl}@med.uni-rostock.de

#### Content:

1. Validation of the microarray data by real-time PCR
2. Performance evaluation of the TILAR algorithm

## 1 Validation of the microarray data by real-time PCR

The analysis of the microarray data revealed 121 genes with significant expression changes in response to intramuscular IFN- $\beta$ -1a therapy (supplemental table). Of these, 12 were selected to be remeasured by real-time PCR. Six of the genes were down-regulated and six were up-regulated in the microarray data set. Apart from that, 3 additional genes were analyzed by real-time PCR: MX1, B2M and IFIT1. The mRNA and protein levels of MX1 [1,2] and B2M [3,4] are known markers of the biological activity of IFN- $\beta$ . Their maximum increase in expression is reached within 12-36 hours of drug administration. However, this increase is not sustained for a full week (see discussion in the main text). MX1 encodes a protein with antiviral activities, while B2M is a component of major histocompatibility complex (MHC) class I molecules. IFIT1 was included as a further prominent IFN- $\beta$ -inducible gene [5]. Besides, the GAPDH gene expression was used as a housekeeping control. We analyzed the samples obtained before and one week after start of IFN- $\beta$ -1a i.m. therapy for a subset of 14 patients (Pat01-04, Pat07-08, Pat15-21 and Pat24).

Real-time PCR measurements were performed with TaqMan assay reagents according to the manufacturer's instructions on a 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA, USA). The predesigned primers and TaqMan probes were purchased from Applied Biosystems (table S1). Total RNA (1  $\mu$ g) from each sample was reverse transcribed to cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA, USA). The real-time PCR reactions were performed in triplicates and the sample pair of each patient was always processed in a single batch to minimize variability. Fluorescence was screened during each denaturation/extension cycle. The cycle at which the fluorescence from a sample crosses the threshold for detection above background (Ct value) was computed automatically using the SDS 2.3 software (Applied Biosystems, Foster City, CA, USA).

Ct values are inversely proportional to the log of the initial mRNA copy number. Therefore, we preprocessed the data in a way to obtain expression levels that are in a linear relationship with the microarray data. First, we calculated the median Ct value of each triplicate (to reject outliers) and normalized the data of each sample to the median Ct value of GAPDH mRNA in the same sample as follows:  $\Delta Ct_{gene} = Ct_{gene} - Ct_{GAPDH}$ . Second, we transformed the data using the equation  $2^{-\Delta Ct}$ , and scaled the result by a factor of 1000 for convenience. The preprocessed transcript levels were then utilized to compute paired *t*-tests comparing each gene's expression at one week into therapy versus baseline. We repeated this analysis on the microarray data using only the 14 MS patients whose

**Table S1:** Genes examined by real-time PCR

Official Symbol	GeneCards ID	Entrez Gene ID	TaqMan Assay ID
FCER1A	GC01P157526	2205	Hs00758600_m1
TNFSF10	GC03M173706	8743	Hs00234356_m1
CSF1R	GC05M149413	1436	Hs00911250_m1
CDKN1C	GC11M002861	1028	Hs00175938_m1
OAS1	GC12P111807	4938	Hs00973637_m1
AQP9	GC15P056217	366	Hs00175573_m1
ALOX12	GC17P006840	239	Hs00167524_m1
GNAZ	GC22P021728	2781	Hs00157731_m1
ESAM	GC11M124128	90952	Hs00332781_m1
MS4A7	GC11P059902	58475	Hs00960227_m1
CTSW	GC11P065405	1521	Hs00175160_m1
FGFBP2	GC04M015571	83888	Hs00230605_m1
MX1	GC21P041720	4599	Hs00182073_m1
IFIT1	GC10P091142	3434	Hs01675197_m1
B2M	GC15P042790	567	Hs00984230_m1
GAPDH	GC12P006514	2597	Hs99999905_m1

Real-time PCR was performed for a total of 16 genes, including GAPDH as a normalization control. The rightmost column provides the IDs of the TaqMan assays used.

expression was studied via real-time PCR. Moreover, Pearson correlation coefficients  $r$  were calculated for each gene to assess whether the data generated by both techniques correlate. Expression differences and correlations with  $P$ -values below 0.05 were considered statistically significant.

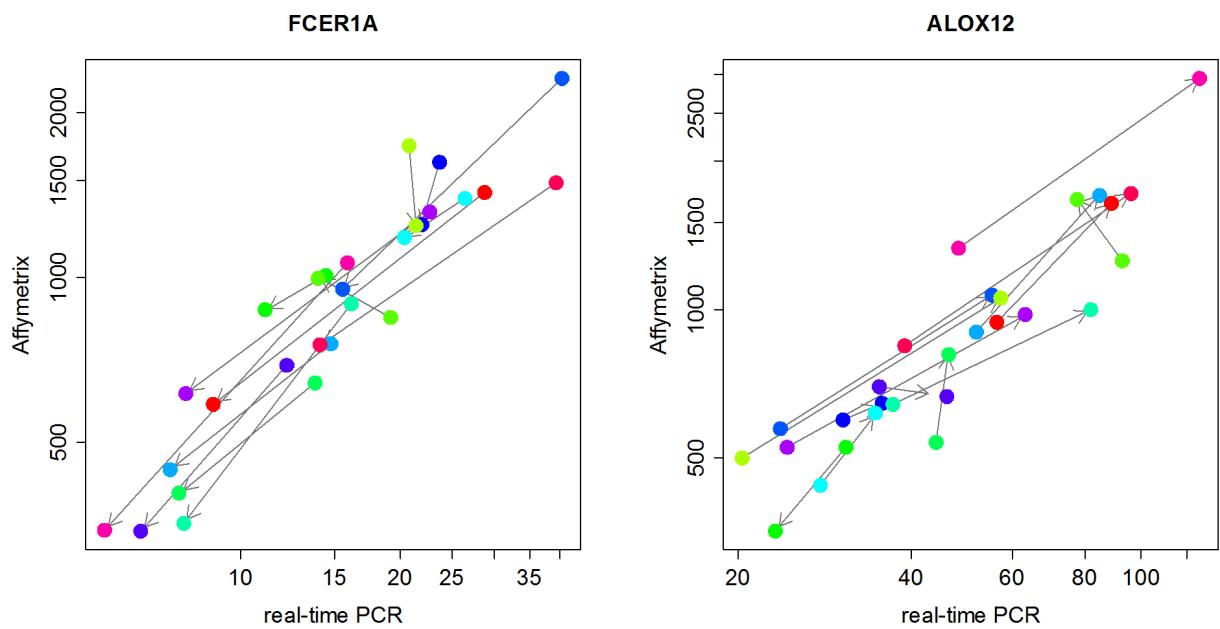
Overall, real-time PCR confirmed the results from the array experiments (table S2). The mRNA levels obtained by the two methods correlated significantly for all genes. Moreover, the real-time PCR data validated the expression differences observed during first week of IFN- $\beta$  treatment for all selected genes. Ten genes were remeasured that were identified as up- or down-regulated after one week in the full microarray data by use of the MAID filtering procedure. Real-time PCR confirmed the significance of these expression changes. The data of two genes, Fc receptor FCER1A and lipoxygenase ALOX12, are shown in figure S1. The results presented here demonstrate that the microarray experiments were performed and analyzed adequately.

**Table S2:** Comparison of the measurements by real-time PCR and Affymetrix microarrays

Official Symbol	Affymetrix data 24 patients (see supplemental table for details)		real-time PCR data 14 patients			Affymetrix data 14 patients			Correlation coefficient (real-time PCR vs Affymetrix)			
	Regulation week 1	Regulation week 4	Mean baseline	Mean week 1	T-Test P week 1 vs baseline	Regulation week 1	Mean baseline	Mean week 1	T-Test P week 1 vs baseline	Regulation week 1	Pearson R	Pearson P
FCER1A			27	15	<0.001		1227	738	<0.001		0.90	<0.001
TNFSF10			304	139	0.007		3637	2665	0.012		0.53	0.004
CSF1R			251	107	0.008		4707	2752	<0.001		0.54	0.003
CDKN1C			106	47	0.005		855	385	<0.001		0.74	<0.001
OAS1			237	219	0.527		1156	791	0.005		-0.09	0.632
AQP9			70	108	0.042		1690	2475	0.027		0.89	<0.001
ALOX12			96	122	0.022		737	1207	0.002		0.78	<0.001
GNAZ			187	234	0.332		873	1282	0.007		0.65	<0.001
ESAM			151	188	0.456		356	739	0.004		0.40	0.037
MS4A7			92	39	0.002		2808	1422	<0.001		0.53	0.004
CTSW			58	69	0.124		2687	3094	0.040		0.69	<0.001
FGFBP2			80	112	0.015		5993	7071	0.132		0.84	<0.001
MX1			206	151	0.366		2178	1561	0.219		0.97	<0.001
IFIT1			5	3	0.448		667	464	0.199		0.68	<0.001
B2M			17131	17399	0.779		24768	25504	0.172		0.43	0.024

Down-regulated  
 P-value < 0.05

We examined 12 genes with significant expression changes during first month of IFN- $\beta$ -1a i.m. therapy according to the array data (upper part) and 3 additional genes (lower part) by real-time PCR analysis. On the basis of the real-time PCR and corresponding microarray data of a subgroup of 14 patients each gene was retested for differential expression between week one and baseline. Cells in the columns "Regulation week 1" visualize whether a gene was found significantly up- or down-regulated or not. The two rightmost columns provide Pearson correlation coefficients  $r$  and the respective  $P$ -values for evaluating whether the real-time PCR data resemble the array data.



**Figure S1:** Gene expression levels obtained by real-time PCR (x-axis) and DNA microarrays (y-axis) for two selected genes. FCER1A (left) was found significantly down-regulated, and ALOX12 (right) was found significantly up-regulated consistently in both measurements. There are two data points for each patient, displaying the expression before first and second IFN- $\beta$  injection, respectively. The data pairs are drawn in a different color for each patient and are linked by an arrow (always pointing to the one week sample). Despite a good correlation of the two types of data, the interindividual differences at the transcript level become evident.

## 2 Performance evaluation of the TILAR algorithm

An objective assessment of the global performance of a GRN inference method is challenging, because there is no "gold-standard" network whose detailed interactions are perfectly known (or at least known with high confidence) [6]. All we have today is a rather fragmentary and only partially correct perception of the (regulatory) interactions between genes. In our original paper on TILAR, time-course Affymetrix microarray data of 19 rheumatoid arthritis patients were used for GRN reconstruction. We then evaluated TILAR's performance by comparing the inferred network (or rather the path of models from an empty to a fully connected network) with gene interactions that were automatically extracted from the literature by text-mining [7]. Such text-mining information is expected to be incomplete and contain errors, and thus might be better referred to as "bronze standard". As a result TILAR outperformed the other algorithms tested and best inference quality was achieved with the adaptive TILAR approach. The latter utilizes gene expression data and TFBS predictions, and furthermore incorporates prior knowledge on gene-TF interactions by introducing additional weights  $\delta$  on the parameters in the model. The lower the weights  $\delta_j$  for the parameters  $\beta_j$  that represent known or suspected gene-TF interactions, the more these interactions are preferred to be in the final model (default is  $\delta_j=1.0$  for no gene-TF interaction preference). Here, the same evaluation procedure was applied, but in comparison to the original publication a different data set (the microarray data of the 24 MS patients) is used for reverse engineering the regulation structure of a different set of genes.

First, we selected genes that are most frequently co-mentioned in the context of MS in PubMed. A respective list of genes was obtained from the Autoimmune Disease Database (version 1.2 as of October 22, 2009), which is a literature-based database that provides gene-disease associations of all autoimmune diseases [8]. Out of the top 50 genes cataloged for "multiple sclerosis", 42 genes were measured in the Affymetrix data set (table S3). We then applied the software PathwayArchitect 2.0.1 to obtain literature-derived gene-gene interactions between them. As most of the 42 genes are well studied, many gene regulatory relationships were found, so that a reliable GRN inference assessment is possible. In total, we retrieved 303 (undirected) gene-gene interactions of type "expression" or "regulation" as labeled by PathwayArchitect. TILAR constrains genes to regulate other genes via one or more TFs. Gene-gene interactions are thus only implicitly defined in the inferred network by gene-TF and TF-gene interactions. Therefore, the knowledge used to assess the method is different from the knowledge integrated during the modeling. Apart

**Table S3:** List of the 50 most frequently mentioned genes in the context of MS

Official Symbol	#PMIDs	GeneCards ID	Entrez Gene ID	Official Full Name	Mean	SD
MBP	1181	GC18M072819	4155	myelin basic protein	576	156.02
TNF	820	GC06P031652	7124	tumor necrosis factor (TNF superfamily, member 2)	412	120.44
IFNG	813	GC12M066834	3458	interferon, gamma	237	67.67
IFNB1	806	GC09M021067	3456	interferon, beta 1, fibroblast	15	8.64
IL10	360	GC01M205007	3586	interleukin 10	85	43.15
IL4	303	GC05P132037	3565	interleukin 4	24	16.86
MOG	291	GC06P029732	4340	myelin oligodendrocyte glycoprotein	48	20.72
TRAT1	281	GC03P110026	50852	T cell receptor associated transmembrane adaptor 1	666	219.86
IL2	268	GC04M123652	3558	interleukin 2	22	16.05
IL6	212	GC07P022732	3569	interleukin 6 (interferon, beta 2)	100	29.61
TGFB1	164	GC19M046528	7040	transforming growth factor, beta 1	1441	373.28
IL1B	143	GC02M113303	3553	interleukin 1, beta	617	155.16
ICAM1	142	GC19P010247	3383	intercellular adhesion molecule 1	494	92.41
CD4	136	GC12P006769	920	CD4 molecule	683	144.07
IFNA1	128	—	—	—	—	—
HLA-DRB1	123	—	—	—	—	—
NOS2A	113	—	—	—	—	—
GFAP	101	GC17M040338	2670	glial fibrillary acidic protein	50	35.92
MMP9	93	GC20P044070	4318	matrix metalloproteinase 9 (gelatinase B, 92kDa gelatinase)	1164	870.68
CD8A	85	GC02M086865	925	CD8a molecule	2248	823.61
VCAM1	84	GC01P100897	7412	vascular cell adhesion molecule 1	94	44.20
CCL5	83	GC17M031222	6352	chemokine (C-C motif) ligand 5	8054	2168.19
HLA-B	81	—	—	—	—	—
HLA-DQB1	72	GC06M032735	3119	major histocompatibility complex, class II, DQ beta 1	867	237.92
CCL2	68	GC17P029606	6347	chemokine (C-C motif) ligand 2	76	298.61
MICB	67	—	—	—	—	—
APOE	64	GC19P050100	348	apolipoprotein E	18	10.81
MAG	64	GC19P040474	4099	myelin associated glycoprotein	17	15.83
CCR5	63	—	—	—	—	—
PLP1	61	GC0XP102837	5354	proteolipid protein 1	48	32.07
NFKB1	60	GC04P103641	4790	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	1921	281.71
CD86	59	GC03P123256	942	CD86 molecule	837	163.03
CTLA4	59	GC02P204440	1493	cytotoxic T-lymphocyte-associated protein 4	163	56.79
CCL3	57	—	—	—	—	—
PER1	57	GC17M007984	5187	period homolog 1 ( <i>Drosophila</i> )	422	147.35
BTNL2	55	GC06M032470	56244	butyrophilin-like 2 (MHC class II associated)	24	9.51
ITGA4	55	GC02P182029	3676	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)	1077	350.58
CXCR3	53	GC0XM070752	2833	chemokine (C-X-C motif) receptor 3	173	67.53
POMC	53	GC02M025295	5443	proopiomelanocortin	29	12.89
ALB	52	GC04P074509	213	albumin	47	22.84
IL2RA	50	GC10M006093	3559	interleukin 2 receptor, alpha	205	57.80
CD40	49	GC20P044181	958	CD40 molecule, TNF receptor superfamily member 5	369	53.13
FAS	48	GC10P090741	355	Fas (TNF receptor superfamily, member 6)	644	142.17
LTA	48	—	—	—	—	—
OMG	48	GC17M026645	4974	oligodendrocyte myelin glycoprotein	173	61.32
CD28	47	GC02P204279	940	CD28 molecule	287	75.25
CNP	47	GC17P037372	1267	2',3'-cyclic nucleotide 3' phosphodiesterase	823	211.84
IL5	46	GC05M131905	3567	interleukin 5 (colony-stimulating factor, eosinophil)	35	23.81
PROM1	46	GC04M015578	8842	prominin 1	51	29.13
FASLG	45	GC01P170894	356	Fas ligand (TNF superfamily, member 6)	410	104.41

The column "#PMIDs" shows (in descending order) the number of PubMed abstracts where the gene occurs in the context of MS. These numbers were derived from the Autoimmune Disease Database. A subset of 42 genes was contained in the microarray data and used to evaluate the performance of the TILAR network inference approach. For each measured gene the corresponding "GeneCards ID" is given as well as mean and standard deviation (SD) over the 72 samples in the data set.

from that, the gene-gene interactions provided by PathwayArchitect also allow to benchmark other inference techniques which do not employ other information in addition to gene expression data.

For constructing a GRN of the 42 genes with TILAR, we analyzed their regulatory regions and found DNA-binding sites of 8 TF entities (table S4) overrepresented at the significance level  $\alpha=0.1$ . These TFs are connected to 28 out of the 42 genes by means of 56 TF-gene interactions. We then applied TILAR to fit a linear model to the expression data while including only a subset of these predicted TF-gene interactions. The precise model parameters (representing gene-TF interactions) were estimated by least angle regression (LARS). LARS builds up estimates for the parameters (regression coefficients) in successive steps, each step adding one covariate to the model, so that gradually all parameters are set non-zero. In simple terms, LARS is a fast and less greedy version of traditional forward selection methods. LARS defines in each step the presence of more and more gene-TF interactions and thus (indirectly) the presence of more and more gene-gene interactions. In this way, LARS specifies network models with different degrees of network connectivity.

Next, we tested whether the inferred edges between genes do exist in the literature-derived network of 303 gene-gene interactions. For this purpose, we calculated the evaluation metrics recall, precision and false positive rate for different network connectivities (i.e. in case of LARS for each regression step). Two common visualizations for assessing how well an inferred network approximates the "gold standard" network are: 1) a plot of the recall versus the precision performance of a method (RPC) and 2) the receiver operating characteristic (ROC) curve, where the false positive rate is plotted against the recall. The "area under the curve" (AU) of both, RPC and ROC, are standard performance measures and have also been applied to evaluate network predictions within the DREAM (Dialogue for Reverse Engineering Assessments and Methods) project [6]. We computed these AU values by the integral of the curve that results from a linear interpolation between the points in the RPC and ROC plot, respectively. In case of the AU(RPC), this yields only an approximation, as here a linear interpolation is not correct [6]. Though, we used the hyperbolic and more accurate functional form at the end of the recall-precision trajectory in order to adequately account for incomplete network reconstructions, that is, if the GRN inference method does not include some edges at all. To assess the statistical significance of a particular AU value, we computed the AU(ROC) and the AU(RPC) for 1000 random GRN predictions (RAND). The RAND algorithm randomly assigns interactions between genes until a fully connected network results. We then calculated the probability (*P*-value) that an AU value obtained by RAND will be higher than the AU value of TILAR (or another method). The *P*-values were computed by 1 minus

**Table S4:** TFBS overrepresented in the regulatory regions of the selected genes

TF Symbol	Transfac Accession	Official Full Name	P-value	Expected Count	Count
TBP	M00216, M00252, M00471	TATA box binding protein	0.001	2.72	9
RUNX1	M00271	runt-related transcription factor 1	0.005	2.70	8
NFATC1, NFATC2, NFATC3, NFATC4	M00302	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1-4	0.005	1.13	5
MEF2A	M00006, M00231, M00232, M00233, M00026	myocyte enhancer factor 2A	0.011	3.12	8
SOX5, SOX9, SRY	M00042, M00160, M00410	SRY (sex determining region Y)-box 5 and 9, sex determining region Y	0.050	3.41	7
IRF1, IRF2	M00062, M00063	interferon regulatory factor 1 and 2	0.063	2.16	5
NFKB1, NFKB2, REL, RELA	M00051, M00052, M00053, M00054, M00194, M00208	nuclear factor of kappa light polypeptide gene enhancer in B-cells family members	0.067	5.99	10
HNF1A	M00132, M00206	HNF1 homeobox A	0.068	1.55	4
<b><math>\Sigma = 56</math></b>					

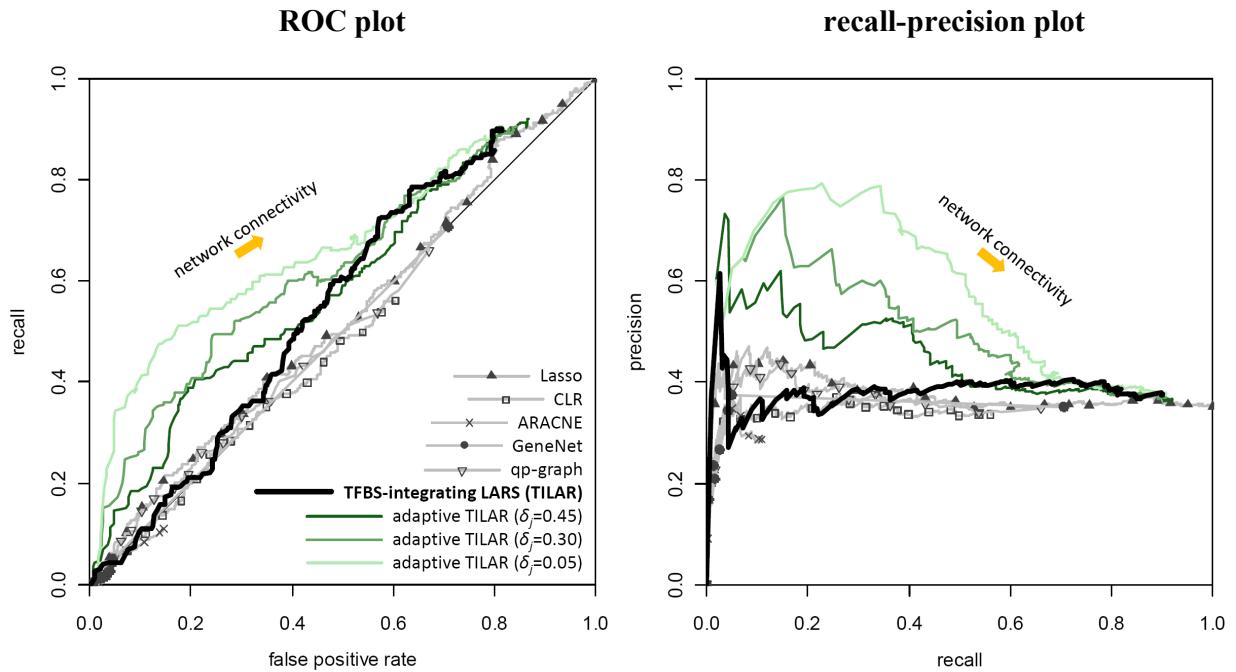
There were 8 consolidated TFs whose DNA-binding sites occurred more often in the regulatory region of the 42 genes than expected by chance ( $P\text{-value}<0.1$ ).

the cumulative probability, evaluated at the AU value of the respective inference method, of the normal distribution having the mean and standard deviation of the RAND AU values. The accuracy of those empirical  $P$ -values is limited due to the rather small number of RAND iterations and the fact that a normal distribution only approximates the AU value distribution in the interval [0,1], but for presented benchmarking analysis this should suffice.

In addition to the assessment of inference qualities of TILAR and adaptive TILAR, we also compared the results of 5 other GRN inference methods: Lasso [9], CLR [10], ARACNE [11], GeneNet [12], and qp-graph [13]. The Lasso (least absolute shrinkage and selection operator) is implemented by LARS and, in this case, aims to find a sparse solution that expresses each gene's expression as a linear function of the expression of other genes. In comparison, TILAR includes predicted TF-gene interactions by the design of the system of linear equations and describes gene-gene interactions indirectly through connections between TFs and genes. Hence, the only difference between the Lasso approach and TILAR is, in fact, the additional constraint in TILAR that the regulatory effect of a gene via a particular TF is the same for each target gene of this TF. CLR and ARACNE use mutual information to describe gene regulatory relationships, and GeneNet and qp-graph compute a partial correlation network. These algorithms build (undirected) gene association networks (or rather networks with different sparsity dependent on the cut-off for inferred edge weights) and have been implemented using the R packages minet, GeneNet and

qpgraph. The entire set of Lasso solutions was computed by the R package lars. All methods were run with default settings and applied on the same gene expression data set. The data were standardized, i.e. the expression levels were transformed for each of the 42 selected genes so that they have mean 0 and variance 1. The adaptive variant of TILAR was used to integrate prior knowledge on gene-TF interactions. A gene-TF interaction describes a gene as a putative regulator of a TF's activity. For instance, a gene may encode proteins that regulate the mRNA level of a particular TF or post-translationally modify the TF, affect its cellular localization, compete for its DNA-binding sites or participate in its upstream signaling pathways. In consequence, as gene-TF edges have no straightforward physical interpretation, their nature is rather phenomenological and accurate knowledge on gene-TF edges is difficult to obtain. Here, we used PathwayArchitect 2.0.1 to retrieve potential (literature-derived) influences of genes on TFs of type "expression", "regulation" or "protein modification". In this way, we found 34 (directed) interactions between the 42 genes and 8 TF entities. We then applied the adaptive TILAR method with three different weights for the coefficients of preferred edges ( $\delta_j=0.45$ ,  $\delta_j=0.30$  and  $\delta_j=0.05$ ). Apart from that, we implemented another random algorithm (gene-TF-RAND) which, similar to TILAR, constructs regulatory networks of TF-gene and gene-TF interactions. Gene-TF-RAND does not consider the microarray gene expression data, but includes all the 56 predicted TF-gene interactions. The method represents a "use TFBS only" inference strategy as gene-TF edges are just randomly added to the network. As for TILAR, gene-TF interactions were not allowed when gene and TF were already connected by a TF-gene interaction, and gene-gene interactions result implicitly. The AU values of gene-TF-RAND were calculated by the mean of 1000 repeated runs.

Figure S2 shows the RPC and ROC curves that resulted by evaluating the global performance of the different inference algorithms using the 303 (regulatory) gene-gene interactions from PathwayArchitect. It is easy to see that TILAR and adaptive TILAR outperformed all other modeling algorithms. The maximum precision of TILAR is 0.615 and of adaptive TILAR 0.793. In comparison, the precision of all the other methods reaches maximum 0.472 (qp-graph). AU values and the respective *P*-values that indicate whether the algorithm's AU values are higher than for RAND are given in table S5. The table shows that the interactions inferred by TILAR match the interactions derived from literature significantly more than random, whereas this was not the case for CLR, ARACNE, GeneNet, Lasso and qp-graph. Hence, the overall performance of the methods, which do not integrate different types of biological information, is poor. One reason for this is that (despite mathematical differences) they all infer interactions solely based on gene co-expression. As



**Figure S2:** ROC and recall-precision curves of different GRN modeling strategies. The better a method performs, the closer its ROC and RPC curve will be to the upper-right and upper-left corner, respectively. TILAR outperformed CLR, ARACNE, GeneNet, Lasso and qp-graph in inferring the regulatory relationships between the 42 selected genes. If adequate prior knowledge on gene-TF interactions is available, the adaptive variant of TILAR can be used to further increase the inference quality.

was discussed by Zampieri et al. for *E. coli* and yeast, such co-expression patterns tend to unveil stable functional categories (e.g. co-participation in a protein complex, genomic co-localization or similar biological function) rather than transient or condition-specific interaction structures (e.g. TF-DNA binding) [14]. Moreover, when trying to infer the regulator from a group of co-expressed genes, there is a risk to arbitrarily select one of these genes. The constraints realized by TILAR incorporate TFBS (and literature) information and thus aid in detecting true regulatory gene interactions more reliably. Interestingly, gene-TF-RAND also showed relatively high AU values. This suggests that TILAR performs well because of both the quality of TFBS predictions and data-fitting using LARS. Additional knowledge on potential gene-TF interactions was integrated by the adaptive TILAR. Dependent on the parameters  $\delta_j$  that express the confidence in this prior knowledge the inference quality increased considerably. However, a lower  $\delta_j$  than 0.05 did not improve the result much.

**Table S5:** Benchmark result for TILAR and other GRN inference algorithms

Network Inference Method	AU(ROC)	AU(ROC) P-value	AU(RPC)	AU(RPC) P-value
TILAR $\delta_j=0.05$	0.68	<2.22 E-16	0.57	<2.22 E-16
TILAR $\delta_j=0.30$	0.63	3.78 E-11	0.50	<2.22 E-16
TILAR $\delta_j=0.45$	0.59	2.07 E-06	0.45	3.60 E-09
TILAR $\delta_j=1.00$	0.56	0.001	0.38	0.064
gene-TF-RAND	0.55	0.010	0.38	0.046
Lasso	0.53	0.095	0.38	0.085
qp-graph	0.50	0.414	0.36	0.363
GeneNet	0.50	0.491	0.35	0.487
CLR	0.48	0.779	0.34	0.782
ARACNE	0.48	0.822	0.33	0.904

The rating of the methods is based on the area under the RPC and ROC curves. An AU(ROC) very close to 0.5 is no better than random. For the adaptive variant of TILAR (the first 3 rows of the table)  $\delta_j$  is set lower than 1.0. *P*-values for the adaptive TILAR with  $\delta_j=0.05$  were below the computational accuracy of R (2.22 E-16). The table clearly demonstrates that the integration of various types of data is much more suited to model gene regulation.

To summarize, we reconfirmed that we can utilize information on TF-gene interactions (predicted by TFBS overrepresentation analysis) and gene-TF interactions (derived from the literature by text-mining) to construct GRN models that describe regulatory interactions between genes (and TFs) with superior accuracy. The benchmarking analysis generated results comparable to our study where we first introduced TILAR [7]. In the original publication on TILAR we discuss the algorithm's performance in more detail and point out remaining issues, e.g. the fact that databases containing TFBS sequence motifs are still by far not complete. We showed that it is important to include additional biological information into the reverse engineering of GRNs. Eventually, the integration of heterogeneous data and prior biological knowledge is increasingly relevant in the whole field of computational biology.

## References

- [1] Gilli, F., F. Marnetto, M. Caldano, A. Sala, S. Malucchi, A. Di Sazio, M. Capobianco, and A. Bertolotto. 2005. Biological responsiveness to first injections of interferon-beta in patients with multiple sclerosis. *J. Neuroimmunol.* 158(1-2): 195-203.
- [2] Stürzebecher, S., R. Maibauer, A. Heuner, K. Beckmann, and B. Aufdembrinke. 1999. Pharmacodynamic comparison of single doses of IFN-beta1a and IFN-beta1b in healthy volunteers. *J. Interferon Cytokine Res.* 19(11): 1257-1264.
- [3] Santos, R., B. Weinstock-Guttman, M. Tamaño-Blanco, D. Badgett, R. Zivadinov, T. Justinger, F. Munschauer 3rd, and M. Ramanathan. 2006. Dynamics of interferon-beta modulated mRNA biomarkers in multiple sclerosis patients with anti-interferon-beta neutralizing antibodies. *J. Neuroimmunol.* 176(1-2): 125-133.
- [4] Buchwalder, P. A., T. Buclin, I. Trinchard, A. Munafò, and J. Biollaz. 2000. Pharmacokinetics and pharmacodynamics of IFN-beta 1a in healthy volunteers. *J. Interferon Cytokine Res.* 20(10): 857-866.
- [5] Goertsches, R. H., M. Hecker, and U. K. Zettl. 2008. Monitoring of multiple sclerosis immunotherapy: from single candidates to biomarker networks. *J. Neurol.* 255 Suppl 6: 48-57.
- [6] Stolovitzky, G., R. J. Prill, and A. Califano. 2009. Lessons from the DREAM2 Challenges. *Ann. N. Y. Acad. Sci.* 1158: 159-195.
- [7] Hecker, M., R. H. Goertsches, R. Engelmann, H. J. Thiesen, and R. Guthke. 2009. Integrative modeling of transcriptional regulation in response to antirheumatic therapy. *BMC Bioinformatics.* 10:262.
- [8] Karopka, T., J. Fluck, H. T. Mevissen, and A. Glass. 2006. The Autoimmune Disease Database: a dynamically compiled literature-derived database. *BMC Bioinformatics.* 7: 325.
- [9] van Someren, E. P., L. F. A. Wessels, M. J. T. Reinders, and E. Backer. 2002. Regularization and noise injection for improving genetic network models. In *Computational And Statistical Approaches To Genomics*. W. Zhang, and I. Shmulevich, eds. Kluwer Academic Publishers, Boston, MA. p. 211-226.
- [10] Faith, J. J., B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5(1): e8.
- [11] Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics.* 7 Suppl 1: S7.

- [12] Opgen-Rhein, R., and K. Strimmer. 2007. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.* 1: 37.
- [13] Castelo, R., and A. Roverato. 2009. Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *J. Comput. Biol.* 16(2): 213-227.
- [14] Zampieri, M., N. Soranzo, D. Bianchini, and C. Altafini. 2008. Origin of co-expression patterns in *E. coli* and *S. cerevisiae* emerging from reverse engineering algorithms. *PLoS One*. 3(8): e2981.

## Danksagung

An dieser Stelle möchte ich mich bei all denen bedanken, die direkt oder indirekt zum Gelingen dieser Arbeit beigetragen haben. Mein besonderer Dank gilt PD Dr. Reinhard Guthke für die Überlassung des Themas und die wissenschaftliche Betreuung der Dissertation. Ich danke ihm für seine stets freundliche Unterstützung dieser Arbeit, für die fachlichen Diskussionen, Kommentare und Anregungen sowie das mir entgegengebrachte Vertrauen und die gewährten Freiräume. Den Großteil meiner Doktorarbeitszeit durfte ich in Rostock verbringen. Daher danke ich Prof. Dr. Hans-Jürgen Thiesen, der mir angenehme Arbeitsbedingungen am Institut für Immunologie der Universität Rostock ermöglicht hat. Ich danke Dr. Peter Lorenz und Dr. Dirk Koczan für die kompetente Beantwortung vieler Fragen und die interessanten Gespräche, die mir wesentliche Impulse für meine Arbeit gegeben haben. Prof. Dr. Uwe Klaus Zettl und Dr. Robert Goertsches gebührt mein Dank dafür, dass sie mich in die Welt der "Multiple Sklerose-Therapien" entführt haben. Ich danke ihnen für die freundschaftliche, konstruktive und oft auch humorvolle Zusammenarbeit und freue mich auf die gemeinsamen bevorstehenden Projekte. Vielen Dank auch an die anderen Rostocker Kollegen, vor allem meinem sehr angenehmen ehemaligen "Bürogenossen" Felix Steinbeck, und an die netten Kollegen der Abteilung Systembiologie/Bioinformatik am HKI in Jena für die moralische Unterstützung und den fachlichen Meinungsaustausch. Ebenso möchte ich all meinen Freunden in der Heimat und anderswo danken, die für Ablenkung und Freude neben der Arbeit gesorgt haben. Insbesondere danke ich Neu-Dr. Robby Engelmann und "Freulein" Wiebke, die entscheidend dazu beigetragen haben, dass ich mich in Rostock stets superwohl gefühlt habe. Was wäre die Arbeit ohne den privaten Ausgleich! Weiterhin danke ich Dr. Jorge Cham, der mich in Phasen der Prokrastination mit seinen Comics daran erinnerte, dass es anderen Doktoranden weltweit ähnlich ergeht. Ein großer Dank geht natürlich auch an meine Eltern und an meinen Bruder. Sie standen mir immer mit Worten und Taten zur Seite.

## Ehrenwörtliche Erklärung

Hiermit erkläre ich ehrenwörtlich, dass mir die geltende Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena bekannt ist. Ich versichere, dass ich die vorliegende Arbeit selbständig und ohne die unzulässige Hilfe Dritter angefertigt habe. Alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen habe ich in meiner Arbeit angegeben. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes haben mich die in der Danksagung genannten Personen unterstützt. Ich habe nicht die Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Arbeit stehen. Weiterhin erkläre ich, dass diese Dissertation noch nicht für eine staatliche oder andere wissenschaftliche Prüfung vorgelegt wurde. Auch habe ich weder diese Dissertation noch eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht.

Jena, den 8. März 2010

Michael Hecker

## Tabellarischer Lebenslauf

**Michael Hecker**  
Diplom-Bioinformatiker

\* 6. April 1982  
in Erfurt

Hans-Sachs-Allee 35  
D-18057 Rostock  
Telefon: +49-162-6505567  
Email: heckinger@freenet.de

**Familienstand:**  
ledig, keine Kinder

**Staatsangehörigkeit:**  
deutsch

Jena, 8. März 2010

## Berufserfahrung

---

- seit Feb. 2010                   **Projektmitarbeiter** des Steinbeis-Transferzentrums  
Proteom-Analyse in Rostock
- Dez. 2007 – Dez. 2008           **Wissenschaftlicher Mitarbeiter** der Gesellschaft  
für Individualisierte Medizin mbH in Rostock
- Nov. 2006 – Jan. 2010           **Doktorand** am Leibniz-Institut für Naturstoff-Forschung  
und Infektionsbiologie e.V. – Hans-Knöll-Institut in Jena

## Studienbegleitende Tätigkeiten

---

- Mai 2005 – Nov. 2006           **Studentische Hilfskraft** am CiS Institut für  
Mikrosensorik gGmbH in Erfurt
- Aug. 2004 – Sep. 2004           **Praktikum** bei Jenoptik Laser, Optik, Systeme GmbH  
in Jena
- Nov. 2003 – März 2004           **Studentische Hilfskraft** an der Friedrich-Schiller-  
Universität Jena

## Studium

---

- Okt. 2006                       **Diplom - Bioinformatiker**  
Friedrich-Schiller-Universität Jena
- Diplomarbeit: „Modellierung der Genregulation von  
Zytokinen während der anti-rheumatischen  
Therapie mit einem TNF- $\alpha$ -Blocker“

## Wehrdienst

---

- Nov. 2000 – Aug. 2001           **Grundwehrdienst** als Richtschütze im  
Panzerbataillon 304 in Heidenheim a.H.

## Schule

---

- Juni 2000                       **Abitur** am Hoffmann-von-Fallersleben Gymnasium  
in Weimar

## Sprachkenntnisse

---

**Englisch** fließend in Wort und Schrift  
**Französisch** - Grundkenntnisse

## Publikationen

---

- Hecker M, Goertsches RH, Fatum C, Koczan D, Thiesen HJ, Guthke R, Zettl UK: Network analysis of transcriptional regulation in response to intramuscular interferon-beta-1a multiple sclerosis treatment. Eingereicht bei *Pharmacogenomics J.* 2010.
- Herlyn P, Müller-Hilke B, Wendt M, Hecker M, Mittlmeier T, Grasl G: Frequencies of polymorphisms in cytokines, neurotransmitters and adrenergic receptors in patients with complex regional pain syndrome type I after distal radial fracture. *Clin J Pain* 2010, 26(3):175-181.
- Goertsches RH, Hecker M, Koczan D, Serrano-Fernández P, Möller S, Thiesen HJ, Zettl UK: Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN-beta-1b treatment in relapsing remitting MS. *Pharmacogenomics* 2010, 11(2):147-161.
- Röwer C, Vissers JP, Koy C, Kipping M, Hecker M, Reimer T, Gerber B, Thiesen HJ, Glocker MO: Towards a proteome signature for invasive ductal breast carcinoma derived from label-free nanoscale LC-MS protein expression profiling of tumorous and glandular tissue. *Anal Bioanal Chem.* 2009, 395(8):2443-2456.
- Hecker M, Goertsches RH, Engelmann R, Thiesen HJ, Guthke R: Integrative modeling of transcriptional regulation in response to antirheumatic therapy. *BMC Bioinformatics* 2009, 10:262.
- Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R: Gene regulatory network inference: data integration in dynamic models - a review. *Biosystems* 2009, 96(1):86-103.
- Goertsches RH, Hecker M, Zettl UK: Monitoring of multiple sclerosis immunotherapy: from single candidates to biomarker networks. *J Neurol.* 2008, 255(Suppl 6):48-57.
- Koczan D, Drynda S, Hecker M, Drynda A, Guthke R, Kekow J, Thiesen HJ: Molecular discrimination of responders and nonresponders to anti-TNF alpha therapy in rheumatoid arthritis by etanercept. *Arthritis Res Ther.* 2008, 10(3):R50.

## Vorträge

---

- Hecker M: Analysing antibody binding patterns by peptide microarrays. Workshop zum "5. Rostocker Proteomforum", 9.-11. Dezember 2009, Rostock.
- Hecker M, Goertsches RH, Thiesen HJ, Zettl UK, Guthke R: Integrative modelling of gene regulatory networks using TILAR. Internationaler Workshop "Integrative Network Inference in Systems Biology", 8.-9. Oktober 2009, Jena.
- Hecker M: Integrative modelling of transcriptional regulation in response to antirheumatic therapy. JCB Workshop, 6. Februar 2009, Jena.
- Hecker M, Goertsches RH, Koczan D, Kekow J, Thiesen HJ, Guthke R: Integrative modelling of gene regulatory interactions relevant in antirheumatic therapy. Internationaler Workshop "Gene Regulatory Network Inference", 25.-26. September 2008, Jena.
- Hecker M, Lorenz P, Kreutzer M, Qian Z, Hong L, Thiesen HJ, Guthke R: Analysing antibody binding patterns by peptide microarrays. Internationaler Workshop "Transcriptome and Proteome Data Analysis and Warehousing towards Systems Biology", 12.-13. Juni 2008, Stuttgart.
- Hecker M: Network inference to understand the responsiveness to antirheumatic therapy. Internationaler Workshop "Data and Knowledge Based Biomolecular Network Reconstruction", 16. März 2007, Jena.
- Hecker M, Drynda S, Kekow J, Thiesen HJ, Guthke R: Dynamic modeling of the response to an autoimmune disease therapy. 2. NiSIS-Symposium, 29. November - 1. Dezember 2006, Puerto de la Cruz, Spanien.