BioSystems 96 (2009) 86–103

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

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

Michael Hecker[a], Sandro Lambeck[a], Susanne Toepfer[b], Eugene van Someren[c], Reinhard Guthke[a,*]

[a] Leibniz Institute for Natural Product Research and Infection Biology - Hans Knoell Institute, Beutenbergstr. 11a, D-07745 Jena, Germany
[b] BioControl Jena GmbH, Wildenbruchstr. 15, D-07745 Jena, Germany
[c] 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.

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 functional 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 (R. Guthke).

0303-2647/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2008.12.004

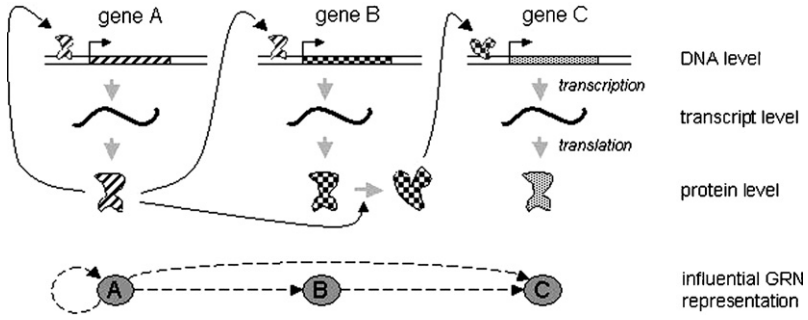

Fig. 1. Schematic view of a simple gene regulatory network. Gene A regulates its own expression and those of gene B. Thereby, gene A might exert its regulatory influence directly (if it encodes a TF) or indirectly (if it controls the activity of another TF possibly via a signalling cascade). When reconstructing 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.

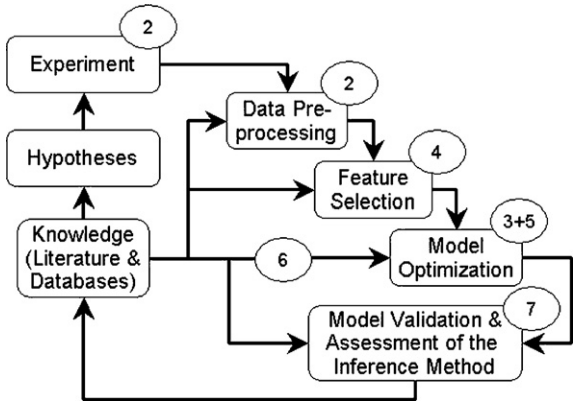

Fig. 2. The systems biology cycle. In this cycle, knowledge leads to new hypotheses, which leads to new experiments, which leads to new models, which leads to new knowledge, etc. Numbers refer to the corresponding sections.

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

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 (Pandey 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 Tegner 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

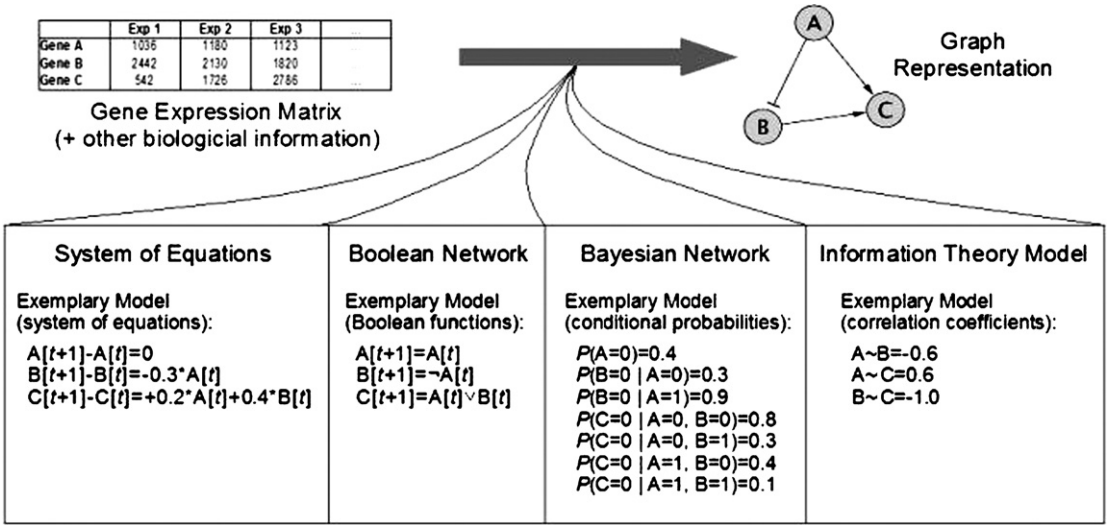

Fig. 3. Exemplary overview of the four main GRN modelling architectures. Here, the aim is to infer the regulatory interactions of three genes (the GRN graph on the top-right) based on the expression data of these three genes for a handful of experiments (the gene expression matrix on the top-left). For 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):

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

where $x(t) = (x_1(t), \ldots, x_n(t))$ is the gene expression vector of the genes $1, \ldots, n$ 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:

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

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):

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

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, α_i and β_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.

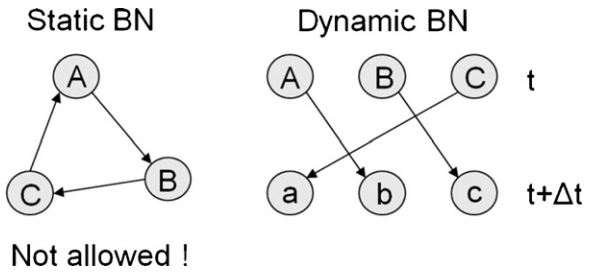

Fig. 4. Difference between static BNs (left panel) and dynamic BNs (right panel). A feedback loop from gene A to gene B to gene C and back to gene A is not allowed in static BNs. However, this feedback loop can be represented in a dynamic BN.

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. ceresivisiae* 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

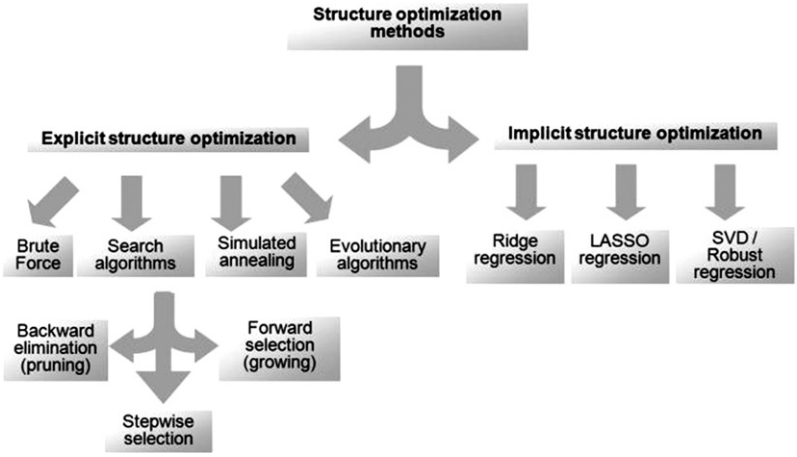

Fig. 5. Overview of structure optimization methods that can be applied in particular for the learning of linear differential/difference equation models.

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 Polisetty 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 $k = 1, 2, 3\ldots$ 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 γ is a constant. In these, so-called scale-free networks, most of the genes are sparsely connected, while a few are very high 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 γ 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. Predefining 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 θ given data D and background knowledge ξ 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 θ, 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 ξ 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 $\sum_j |\beta_j| \leq s, \quad s > 0$. Because of the nature of this constraint it tends to produce some coefficients β_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: $\sum_j \tilde{w}_j |\beta_j| \leq s, \quad s > 0$. Thereby, a relatively low weight $\tilde{w}_j$ provokes that the edge corresponding to β_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 β_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 − 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 where 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	
D'haeseleer et al. (1999)	RTQ-PCR	Rat	65	Time-series (28 tp.)	–	–	65	Linear difference equations	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	√			–
Imoto 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)	√			–
Nariai 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	√			–
Guthke 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.	*T. thermophilus*	612	Time-series (14 tp.)	–	Hierarchical clustering	25	S-system model	Evolutionary algorithm	–	√			–
Bonneau et al. (2006)	Oligon. ma.	*Halobacterium*	~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.)	Fold-criterion	–	101	Linear difference equations	LASSO	–	√		√	External (literature)
Faith et al. (2007)	Oligon. ma.	*E. coli*	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.

Khatri, 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-TNFalpha 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., Basso, 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: Stolla, S.A., Leen, T.K., Muller, 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.
Polisetty, 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.
Spieth, 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.
Spieth, 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.
Tegner, 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., Reinders, M., Backer, E., 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., Reinders, M., 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., Tegner, 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.