EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

Review article:

DATA- AND KNOWLEDGE-BASED MODELING OF GENE REGULATORY NETWORKS: AN UPDATE

Jörg Linde[1], Sylvie Schulze[1], Sebastian G. Henkel[2], Reinhard Guthke[1*]

[1] Research Group Systems Biology / Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute, Beutenbergstr. 11a, 07745 Jena, Germany
[2] BioControl Jena GmbH, Wildenbruchstr. 15, 07745 Jena, Germany

[*] Corresponding author: Reinhard Guthke, Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute, Beutenbergstr. 11a, 07745 Jena, Germany; Tel.: +49 5321083; Fax: +49 5320803; E-mail: reinhard.guthke@hki-jena.de

http://dx.doi.org/10.17179/excli2015-168

ABSTRACT

Gene regulatory network inference is a systems biology approach which predicts interactions between genes with the help of high-throughput data. In this review, we present current and updated network inference methods focusing on novel techniques for data acquisition, network inference assessment, network inference for interacting species and the integration of prior knowledge. After the advance of Next-Generation-Sequencing of cDNAs derived from RNA samples (RNA-Seq) we discuss in detail its application to network inference. Furthermore, we present progress for large-scale or even full-genomic network inference as well as for small-scale condensed network inference and review advances in the evaluation of network inference methods by crowdsourcing. Finally, we reflect the current availability of data and prior knowledge sources and give an outlook for the inference of gene regulatory networks that reflect interacting species, in particular pathogen-host interactions.

Keywords: gene regulatory networks, modeling, reverse engineering, network inference, prior knowledge, RNA-Seq

INTRODUCTION

One characteristic of life is that living organisms constantly adapt to environmental changes (Koshland, 2002). Higher organisms may use their brain for long term reactions or reflexes as nearly instant reactions in response to stimuli. On a molecular basis, microorganisms as well as cells and tissues of higher organisms sense environmental changes (Groisman and Mouslim, 2006). This information is then transmitted into the cells and processed which finally leads to a reaction of the cells. Sensing, transmitting and processing of the information is performed by complex molecular interactions (Miller and Bassler, 2001). While our knowledge of these interactions is still limited, it is obvious that errors in information processing may lead to diseases (Follo et al., 2015; Compston and Coles, 2008; Wang et al., 2012; Glocker et al., 2006).

Systems biology is a research area which aims to understand living systems as a whole, instead of focusing on single biological entities (Ideker et al., 2001). Systems biology is often (but not exclusively) connected with *omics*. Here, researchers characterize and/or quantify (nearly) all biological molecules of a specific type which allows us to

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

study the complete picture of the system. For example, transcriptomics measures the abundance of all transcripts in a sample, while proteomics measures the abundance of all proteins. One way of describing biological systems are networks, i.e. graphical representations, where the nodes represent objects of interest and edges represent relations between these objects (Le Novère et al., 2009). Network models do not only help to explain, understand and describe the functioning of a cell (Barabási and Oltvai, 2004), but also to understand disease progression and to discover drugs (Butcher et al., 2004). In gene regulatory networks (GRNs) nodes are genes and edges represent interactions between genes, such as activation or repression. Genes do not necessarily interact directly with each other. In fact, the most direct interaction is a gene coding for a transcription factor (TF) which binds to the promoter region of another gene and regulates its expression. A gene can also influence the expression of another gene more indirectly, via signaling cascades or whole pathways. For sake of simplicity these influences are also described with the word ‘interaction’.

As the underlying structure of many networks is not (completely) known, one focus of systems biology is uncovering the complex and dynamic interactions between genes (Hecker et al., 2009a). The research area called 'network inference (NI)' aims at the deduction of network structures utilizing high-throughput data with help of reverse engineering techniques. In most cases transcriptome data is used. NI consists of three parts: the identification of potential regulators, the prediction of target genes and the inference of the mode of interaction (e.g. activation or repression). The number of genes may vary from only two genes to full-genomic networks. A general problem in NI is the high dimensionality (thousands of genes) versus the limited number (tens to hundreds) of samples. Thus, GRN inference is underdetermined (‘curse of dimensionality’) implying that there could be many equivalent (indistinguishable) solutions (networks). Motivated by this fundamental problem, there exists a number of NI approaches, which are compared in outstanding review articles (de Jong, 2002; van Someren et al., 2002; Gardner and Faith, 2005; Bansal et al., 2007, Ay and Arnosti, 2011; Wu and Chan, 2012; Emmert-Streib et al., 2014). In 2009, our group thoroughly reviewed NI approaches with a focus on data integration (Hecker et al., 2009a). In the review on hand, we present an update of the former review with a special focus on novel techniques for data acquisition, NI assessment and NI for interacting species.

In what follows, we give an overview about the main NI approaches focusing on novel and updated methods introduced since 2009. During that time, the NI community has emphasized the inference of large-scale or even full-genomic networks, the integration of additional data and the combination of NI methods. The integration of data from various (omics) experiments and knowledge databases into one general model is an important challenge in systems biology (Gomez-Cabrero et al., 2014). In the field of NI, additional data is often integrated with help of prior-knowledge, i.e. predicted or known interactions based on additional data or knowledge sources. Here, we give an overview of commonly used prior-knowledge sources.

The advance of Next-Generation-Sequencing of cDNAs derived from RNA samples (RNA-Seq) allows to study transcriptomes with a so far unreachable depth and quality (Morin et al., 2008). On the other hand, data pre-processing poses new challenges. Here, we describe a work-flow combining RNA-Seq data analysis with NI (Figure 1). In particular, the advance of RNA-Seq allows researchers to perform transcriptome studies of interacting (micro-) organisms using the same technology without having to separate RNA samples (‘dual RNA-Seq’; Westermann et al., 2012). This allows to predict GRNs of organisms which interact with each other. Special interest is in patho-

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

gen-host interaction networks which we present in this review.

A particular challenge is the evaluation of predicted networks. Advances in the evaluation of NI methods by crowdsourcing within the DREAM initiative are described in this review.

GENE REGULATORY NETWORK INFERENCE

NI aims to determine the structure and parameters of GRNs. Due to various sources of perturbation, biological systems adapt gene expression and their functionality. The main biological processes and components as well as a model network representation are shown in Figure 2. To fulfill complex tasks and functionalities of living systems and to adapt to various perturbations gene expression changes with respect to their amount (concentration, activity) and influences the expression of other genes. The terminus 'Gene Regulatory Network' (GRN) refers to components of transcriptional regulation, i.e. target genes and TF genes, specific products of gene expression allowing for a complex regulatory response. Understanding of GRNs means understanding of underlying mechanisms and the potential for targeted manipulation of biological systems (Figure 2).

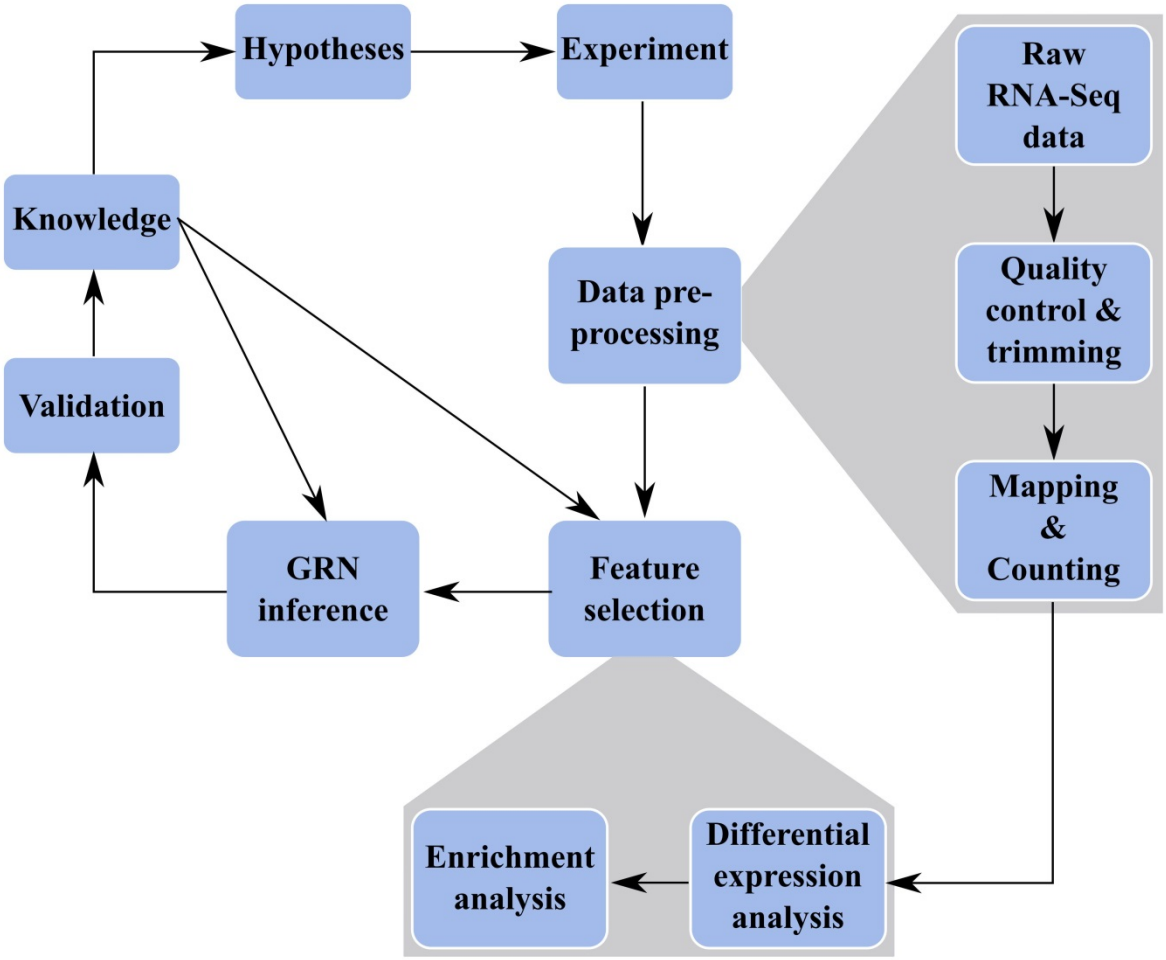

Figure 1: Workflow of GRN inference. Systems Biology Cycle of wet lab (experiment) and dry lab work: Experiments lead to RNA-Seq data, which need to be preprocessed and features have to be selected (more detailed steps are shown in grey boxes). A GRN is inferred for selected features. Predicted interactions are validated leading to more knowledge and new hypotheses. Both analysis of experimental data (data preprocessing and feature selection) and modeling (network inference) is supported by prior knowledge.

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

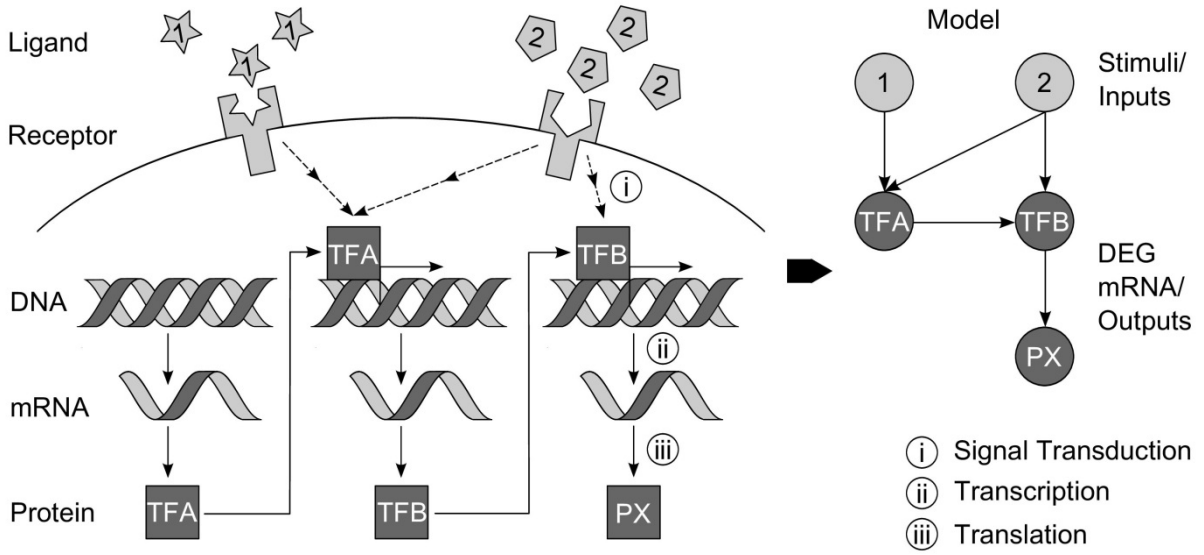

Figure 2: Regulated gene expression and model network representation. External stimuli (ligands binding to receptors on the cell's surface) may trigger an alteration in gene expression. Via signal transduction, the most important regulators, the transcription factors, are influenced. They regulate the transcription of DNA to mRNA, which subsequently is translated to proteins. Those regulated biological processes can be transformed to a network model (inference), whose main nodes represent genes or their products (typically on the level of regulated transcription).

GRNs can be described by mathematical network models built up of nodes describing the components of gene expression, i.e. genes or their products, and edges, i.e. connections between nodes denoting an interacting effect. Nodes and edges together form a network that is primarily defined by its structure, which means existing or non-existing connections between nodes. One important property of GRNs is the sparseness, which refers to the fact that in biological networks much less edges occur than theoretically possible, i.e. in a fully connected network (Leclerc, 2008). One additional specific property is the scale-freeness of GRNs. That means the number of edges per gene follows a power law distribution: Many nodes with a low number of connections and few nodes with a high number of connections ('hubs') exist.

The interactions presented as edges in the graphical presentation of GRNs (Figure 2) may exhibit different strengths of interactions or different molecular mechanisms. For describing those properties mathematical models are needed. Nonzero parameters of these models are represented in the graph. By identifying the nonzero parameters of a model, NI reveals the structure of the system. Additionally, parameter optimization methods are applied to estimate the exact value of the parameters which can be interpreted as the strength of the interaction. To infer underlying GRNs of a biological system, NI methods use data measuring the gene expression intensity and/or the abundance/activity of proteins. In most cases transcriptome data is used. Advanced methods need to know the variance of a measurement (via replicates) and temporal resolution. Different assumptions on underlying processes, prior knowledge about molecular mechanisms, available mathematical methods and many more constraints led to the development of different NI methods. Most of the presented NI methods consider transcriptional regulation, measure mRNA levels which shall reflect the major behavior of regulatory processes and not the mechanisms but only the existence of interactions is the

focus of NI, i.e. relatively large data sets are used for an automatic *top-down* modeling.

RNA-SEQ DATA ANALYSIS FOR GENE REGULATORY NETWORK INFERENCE

Data requirements

NI often relies on transcriptome data. Common technologies for high throughput transcriptome studies are microarrays and RNA-Seq where microarrays have been frequently used. Advantages and disadvantages regarding their applications, costs and time consumption, sensitivity and dynamic range of detection, challenges in data analysis and data storage have been outlined in various publications (Mantione et al., 2014; Malone and Oliver, 2011; Wang et al., 2009). In general, RNA-Seq allows researchers to study transcriptomes with a so far unreachable depth and quality (Morin et al., 2008). Furthermore, RNA-Seq allows to study the transcriptome of non-model organisms, since the expensive design and spotting of arrays is not necessary. In fact, RNA-Seq may be performed even without having the genome sequence at hand (CUFFLINKS (Trapnell et al., 2012), TRINITY (Haas et al., 2013). Obviously, in future there will be more NI approaches based on RNA-Seq data. For this reason, we summaries major steps in RNA-Seq based network inference here.

When studying interacting species it is of interest to monitor the transcriptome of all present species. Microarrays can be utilized, when it is known which species interact and microarrays are available for these species. These limitations do not apply for 'dual RNA-Seq', where RNA samples of two (all) species are sequenced together and transcripts are separated *in silico*. This also prevents a possible transcriptional change caused by the experimental separation of species. In the following we summarize RNA-Seq data properties, RNA-Seq platforms that open new opportunities in GRN inference as well as 'dual RNA-Seq'.

RNA-Seq is a powerful technology for transcriptome profiling, but the understanding of data properties is incomplete and standard protocols for data generation and analysis are lacking. Eukaryotic total RNA consists of ~80 % rRNA, ~15 % tRNA and only a small portion of mRNA (Lodish et al., 2000). To increase the informative output of an RNA-Seq study, samples can be depleted for rRNA or enriched for RNA species of interest (e.g. polyadenylated RNAs) (Sims et al., 2014). Nevertheless, this results in a loss of information and has to be considered with caution especially when dual RNA-Seq is carried out. To mention a single example, let's assume one wants to study the transcriptome of human cells infected with bacteria. In bacteria most mRNAs are not polyadenylated, only when they are tagged for degradation whereas in human cells mRNAs are polyadenylated. Consequently, an enrichment for polyadenylated RNAs would result in a misleading RNA sample (Westermann et al., 2012). Furthermore, Lahens et al. (2014) concluded that rRNA depletion introduces a bias in coverage.

Given a fixed budget, a trade-off between the number of sequenced reads and the number of replicates is needed. Based on mRNA enriched samples of the human cell line MCF7, Liu et al. (2013) found that lower sequencing depth, but a higher number of biological replicates increases the power and accuracy to detect differentially expressed genes. Sequencing a number of reads resulting in more than 10 million mapped reads led to a significantly smaller improvement than generating more replicates instead. Nevertheless, the sequencing depth has to be determined taking the research question into account, because an accurate identification of lowly expressed transcripts requires a sufficient amount of reads (Sims et al., 2014).

Another challenge arises from the nature of the human genome – approximately 50 % is constituted of repetitive elements. Typically, only 70-80 % of short reads map uniquely to the human genome depending on read length and availability of paired-end reads. Methods for assigning ambiguous reads have

been briefly reviewed by Treangen and Salzberg (2011).

Recently, the question whether RNA-Seq is reproducible was addressed. The GEUVADIS consortium (Genetic European Variation in Disease, a European Medical Sequencing Consortium) found small technical variation in samples sequenced with Illumina HiSeq2000 using the exact same protocols ('t Hoen et al., 2013). Within the SEQC/MACQ-III (Sequencing Quality Control) project different sequencing platforms across multiple laboratory sites and analysis pipelines were examined for the detection of differential expression. They found reproducibility when filters for p-value, fold-change and expression level were applied (SEQC/MAQC-III Consortium 2014).

Multiple sequencing platforms have been developed over 10 years of next-generation sequencing research. All second generation platforms, i.e. 454, SOLiD and Illumina, are light-based capturing a fluorescence signal. New approaches, termed 'third generation' are emerging such as the light-based PacBio, pH-based Ion Torrent and current-based Oxford Nanopore. A good overview over second and third generation sequencing platforms is outlined in Liu et al. (2013). One advantage of third generation platforms are steadily increasing read lengths. Applying the PacBio system, read lengths > 1400 bp (Mosher et al., 2014), were reported recently. A first paper about the application of the Oxford Nanopore sequencer reports average read lengths of 5 kb, but concludes that a dramatic decrease of error rates is required (Mikheyev and Tin, 2014). Increasing read lengths of third generation technologies will help to overcome problems such as ambiguous mapping, but second generation platforms are still dominating in application.

RNA-Seq data trimming

Typically, the first step of RNA-Seq data analysis (Figure 1) is clipping of sequencing adapters and removing low quality bases ('trimming'), followed by read mapping and counting. Recently, nine trimming algorithms were evaluated on Illumina RNA-Seq data sets (Del Fabbro et al., 2013). They found comparable performances of all tools (ConDeTri, Cutadapt, ERNE-FILTER, FASTX, PRINSEQ, Sickle, SolexaQA, SolexaQA-BWA, Trimmomatic) applying them to a high quality data set of *Arabidopsis thaliana*. Given a lower quality data set of *Homo sapiens,* SolexaQA performed best in terms of keeping the most reads and aligning a high percentage of them. In comparison, other tools such as FASTX did not show good performance. Some trimming tools have been developed for specific platforms, such as Trimmomatic (Bolger et al., 2014) for Illumina data. Besides, trimming tools include different properties regarding e.g. adapter removal or application to paired-end reads (Jiang et al., 2014).

Read mapping

A very important preprocessing step which has great influence at down-stream analysis is the alignment of reads to the reference genome (mapping). In a recent review, ten alignment tools were evaluated regarding multiple properties such as alignment yield, spliced alignments, mismatches and accuracy (Engström et al., 2013). The aligners MapSplice, GSNAP, GSTRUCT and STAR were evaluated as favorable tools, even though the latter three reported many false exon junctions. It was concluded that TopHat2 is an effective tool, although only 84 % of reads were aligned in comparison to 90 % of reads aligned with MapSplice. Alignment tools like TopHat-Fusion, FusionSeq or SplitSeek have been developed to align reads generated from cancer cells, for which fusion genes caused by rearrangement events (e.g. chromosome breakage and rejoining) are common (Treangen and Salzberg, 2011). When analyzing dual RNA-Seq data, the genomic landscape of interacting species has to be taken into account to determine alignment tool parameters. For example, potential host organisms such as human and mouse have a high percentage of intron-containing genes. On the other hand,

the percentage of intron-containing genes in pathogenic fungi varies a lot, e.g. *Candida glabrata* has ~1.5 % and *Cryptococcus neoformans* has ~97 % intronic genes (Ivashchenko et al., 2009), whereas bacteria do not have introns at all.

Counting

After the mapping step, the number of reads assigned to a feature (e.g. exon, transcript, gene) has to be counted to estimate the expression level. This data needs to be corrected for biases, but standard approaches usually depend on the feature length and the non-uniform distribution of reads to features. The new counting approach maxcounts is based on the maximum of the per-base counts and claims to reduce biases in RNA-Seq data (Finotello et al., 2014). So far, maxcounts is only applicable to exon-level. Usually, counting on transcript- or gene-level is of interest for GRN inference for which the very fast featureCounts tool can be applied (Liao et al., 2013).

Differential expression analysis

The main issue of small- and medium-scale GRN inference is the feature selection, i.e. the identification of the 'most important' genes or proteins of interest for a certain system or process. Identification of differentially expressed genes is an important step for feature selection, i.e. to narrow down the number of network nodes.

For RNA-Seq data, various statistical methods have been proposed and recently the performance of the common tools Cuffdiff2, DESeq and edgeR were compared by Zhang et al. (2014). It was recommended to apply DESeq and edgeR preferably to Cuffdiff2, especially when sequencing depth is low (< 10M). The more conservative DESeq detects less differentially expressed genes and shows a lower false positive rate than edgeR. On the other hand, edgeR is more liberal and tolerates unbalanced library sizes and low sequencing depth. Soneson and Delorenzi (2013) evaluated 11 differential expression analysis tools and found similar results regarding DESeq and edgeR. Furthermore, none of the 11 tools performed best under all circumstances and they provide a short overview about the main finding for every tool. Last year, Love et al. (2014) released the updated DESeq2 and compared DESeq2 to six other differential expression analysis tool including DESeq. They found that all algorithms control the false positive rate, whereas DESeq2 is less conservative than DESeq and Cuffdiff2 and more conservative than edgeR, voom and SAMseq.

Gene enrichment analysis

Another option for feature selection is Gene enrichment analysis. Here, genes are grouped based on their function or biological process. The user may identify processes mostly enriched with DEGs and focuses NI only on those genes. In 2009, Huang et al. (2009) listed 68 enrichment tools, classified them in three group and highlighted properties and limitations of each group. Afterwards, more enrichment tools have been published, e.g. KOBAS (Xie et al., 2011). KOBAS 2.0 incorporates knowledge across 1,327 species from 5 pathway databases (KEGG PATHWAY, PID, BioCyc, Reactome and Panther) and 5 human disease databases (OMIM, KEGG DISEASE, FunDO, GAD and NHGRI GWAS Catalog). For other species not or only weakly represented in these databases, there are more specific tools. For instance, the online resource and web tool FungiFun 2.0 was developed for functional analysis of lists of fungal genes and proteins (Priebe et al., 2015). Most enrichment tools assume, that all genes are equally likely to be selected as differentially expressed. Contradictory to this assumption Oshlack and Wakefield (2009) found that long (or highly expressed) transcripts are more likely to be detected as differentially expressed, which also affects enrichment analysis. GOseq is one of the tools that integrates a correction for this selection bias (Young et al., 2010; Rahmatallah et al., 2014).

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

BASIC INFERENCE METHODS

After summarizing how to gain raw data for NI, this chapter gives an overview of basic NI approaches. Please note, that most approaches do not only work on transcriptome data but might also be applied on other (*omics*) data.

Inference algorithms can be classified by their major properties (Table 1). These are (i) the underlying method, (ii) the result, (iii) the directionality of interactions, (iv) the consideration of dynamics and (v) the integration of prior knowledge (PK), i.e. putative or known interactions based on additional data sources such as literature.

(i) The underlying method or framework describes the key aspect that characterizes the inference approach and can be: (a) Boolean modeling, (b) probabilistic modeling, (c) Information theory-based methods (Mutual Information), (d) (linear) regression and (e) (complex) optimization. These main methods will be described in the subsequent paragraphs.

(ii) The result strongly depends on the selected method and can for example be a Bayesian network, a correlation network, a graphical model or a mathematical model consisting of algebraic or differential equations.

(iii) The model network graph may contain directed or undirected edges, i.e. in the former case cause and effect are clearly distinguished while in the latter case there is rather a general relationship.

(iv) Whether the resulting model is dynamic, i.e. the state of the network at a certain time point also depends on its state at former time points) or static mainly depends on the consideration of time series or steady state data.

(v) The integration of PK has been shown to improve the reliability of predicted novel interactions (Hecker et al., 2009a; Greenfield et al., 2013; Isci et al., 2014; Hasegawa et al., 2014; Olsen et al., 2014). Since reliable information about the experimentally verified interaction is increasing, a NI methods integrate PK in different ways most nowadays (see below section '*Integration of prior knowledge*'). One major difference is whether PK is softly integrated, i.e. the NI method may neglect an interaction within the list of PK if it contradicts the measured data and model assumptions.

Further properties of NI methods are the (vi) non-linearity or linearity, (vii) the explicit consideration of stimulation, (vii) the consideration of stochastics and application of probabilities, (viii) the network size, (ix) the number of required data and (x) the availability as a software tool.

(vi) Mathematical models may be linear or non-linear depending on (a) how detailed and realistic molecular mechanisms are described in the model and (b) whether only the behavior in the neighborhood of a perturbed (steady-state) operating point is considered.

(vii) Although the change in gene expression is caused by one or more (external) stimuli only few methods do consider them explicitly.

(viii) The application of probabilities is based on the information that repeated measurements have the property to follow certain distributions originating from stochastic processes of the biological and/or technical system (measurement method). Finally, the network model size strongly correlates with the available data and assumptions that condense that data, e.g. by clustering or focusing on certain pathways.

The basic methods will be described in the following.

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

Table 1: GRN inference methods

Algorithm	Reference	Main Methods	Main Result	Dir.	Dyn.	PK
REVEAL	Liang et al., 1998	MI+BM	Boolean Network	+	+	-
	Akutsu et al., 1999	BM	Boolean Network	+	+	-
	Martin et al., 2007	BM	Boolean Network	+	+	-
	Eduati et al., 2010	BM	Boolean Network	+	+	-
BNT	Murphy and Mian, 1999	PM	Bayesian Network	+	+	+
BANJO	Hartemink et al., 2001	PM	Bayesian Network	+	+	+
DREM	Ernst et al., 2007	PM	Regulatory Graph	-	+	+
ebdbNet	Rau et al., 2010	PM	Bayesian Network	+	+	-
BMA	Yeung et al., 2011	PM	Bayesian Network	+	-	+
iBMA	Lo et al., 2012	PM	Bayesian Network	+	-	+
ScanBMA	Young et al., 2014	PM	Bayesian Network	+	-	+
GeneNet	Schäfer et al., 2006	PC	GGM	+/-	+	-
RELNET	Butte and Kohane, 2000	MI	Correlation Network	-	-	-
ARACNE	Basso et al., 2005	MI	Correlation Network	-	-	+
TD-ARACNE	Zoppoli et al., 2010	MI	Directed Graph	+	+	+
MRNET	Meyer et al., 2007	MI	Correlation Network	-	-	-
CLR	Faith et al., 2007	MI	Correlation Network	-	-	-
tlCLR	Madar et al., 2010	MI	Correlation Network	+	+	-
C3NET	Altay & Emmert-Streib, 2010	MI	Correlation Network	-	-	-
	Küffner et al., 2012	ANOVA	Correlation network	-	-	-
NIR	Gardner et al., 2003	Regression	ODE	+	-	+
MNI	di Bernardo et al., 2005	Regression	ODE	+	-	-
LARS-EN	Zou and Hastie, 2005	Regression	AE	+	-	-
	Gustafsson et al., 2005	Regression	ODE	+	+	-
TSNI	Bansal et al., 2006	Regression	ODE	+	+	-
GENLAB	van Someren et al., 2006	Regression	ODE	+	+	-
Inferelator	Bonneau et al., 2006	Regression	ODE	+	+	-
TlCLR+Inferelator	Greenfield et al., 2010	MI+Regression	Directed Graph	+	+	-
Inferelator+MEN+BBSR	Greenfield et al., 2013	Regression	Directed Graph	+	+	+
TILAR	Hecker et al., 2009b	Regression	AE	+	-	+
exTILAR	Vlaic et al., 2012	Regression	ODE	+	+	+
TIGRESS	Haury et al., 2012	Regression	AE	+	-	-
	Kulkarni et al., 2012	Regression	ODE	+	-	+
gp4grn	Äijö & Lähdesmäki, 2009	PM+Regression	Directed Graph	+	+	-
	Menéndez et al., 2010	PM+Regression	Undirected Graph	-	-	-
	Holter et al., 2001	SVD	ODE	+	+	-
	Yeung et al., 2002	SVD	ODE	+	+	-
GNR	Wang et al., 2006	SVD	ODE	+	+	-
NetGenerator	Guthke et al., 2005	Optimisation	ODE	+	+	+
JCell	Spieth et al., 2006	Optimisation	ODE	+	+	+
	Nelander et al., 2008	Optimisation	ODE	+	+	-
DPLSQ	Nakajima et al., 2012	Optimisation	ODE	+	+	+
	Hasegawa et al., 2014	Optimisation	ODE	+	+	+
	Yip et al., 2010	PM+Opt.	Directed Graph	+	+	-

Table 1 (cont.): GRN inference methods

Algorithm	Reference	Main Methods	Main Result	Dir.	Dyn.	PK
NIMOO	Gupta et al., 2011	MI+Opt.	ODE	+	+	+
	Pinna et al., 2010	GA	Directed Graph	+	-	-
	Flassig et al., 2013	GA	Directed Graph	+	-	+
GENIE3	Huynh-Thu et al., 2010	TBEM	Directed Graph	+	-	+

Abbreviations: TD-ARACNE – TimeDelay-ARACNE; Main categories: MI – Mutual Information; BM – Boolean Modelling; PM – Probabilistic Modelling; PC – Partial Correlation; SVD – Singular Value Decomposition; GA – Graph Analysis; TBEM – Tree-based Ensemble Method; Main results: GGM – Graphical Gaussian Model; ODE – Ordinary Differential Equations; AE – Algebraic Equations; Opt. – Optimisation; Dir – Directed graph; Dyn. – Dynamic model; PK – Prior knowledge integration

Boolean Modeling

At each node (gene) of the dynamic networks resulting from Boolean modeling, the discretized input values are transformed to an output value by Boolean rules (operators) of which AND, OR and NOT are the simplest and most widely used ones. Boolean networks were first introduced by Kauffman (1969), but have later been used as models of GRNs (Akutsu et al., 1999; Martin et al., 2007; Eduati et al., 2010), in particular by the REVEAL algorithm (Liang et al., 1998), a method combining Boolean modeling and information theory elements (see below). The advantage of Boolean networks is their simplicity while drawbacks are the necessity of discretizing the continuous expression values and the limited coverage of real mechanisms given by the Boolean operators.

Probabilistic Modeling / Bayesian Networks

Probabilistic modeling is not restricted to infer GRNs, but is a common class of algorithms in that field. Most popular, probabilistic modeling infers Bayesian Networks (BN) in which the expression of each gene is considered to be a random variable following probability distributions. This major aspect also shows one of the disadvantages: the need for many data to determine the conditional probabilities. Major advantages are their ability to find hidden variables and the common and easy integration of prior knowledge. BNs are typically displayed as directed graphs that can be either static or dynamic. The latter does not follow the concept of feedback but describes the propagation of information from one time step to the next. There exist major toolboxes and packages like BNT (Bayes Net Toolbox) (Murphy and Mian, 1999), BANJO (Hartemink et al., 2001), ebdbNet (Rau et al., 2010) and the family of BMA methods (Bayesian Model Averaging) (Yeung et al., 2011; Lo et al., 2012; Young et al., 2014). Methods of probabilistic modeling are also described by Friedman et al. (2000), Perrin et al. (2003), Markowetz et al. (2005) and Ernst et al. (2007).

Mutual Information / Information Theory Models

Information theory was applied to develop inference methods for GRN by using mutual information. This term measures the statistical dependency between the (discrete) states of two random variables, which represent the expression intensities of two genes. Inference methods then generate correlation-like undirected graphs containing this dependency information. Several algorithms have been developed: RELNET (Butte and Kohane, 2000), ARACNE (Basso et al., 2005), MRNET (Meyer et al., 2007), CLR (Faith et al., 2007) and C3NET (Altay and Emmert-Streib, 2010). Typically, those methods are restricted to static networks, but TimeDelay- ARACNE (Zoppoli et al., 2010) or tlCLR (Madar et al., 2010) are able to generate directed graphs also considering time-dependent information.

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

Linear Models and Regression

Mathematical models can be dynamic or steady state models depending on whether the problem formulation and solution is set up of algebraic or differential (difference) equations. A linear dynamic model can be written as

$$\dot{x}_i = \sum_{j=1}^{N} a_{i,j} x_j + \sum_{k=1}^{M} b_{i,k} u_k \qquad [1]$$

that is the change (temporal deviation) of the *i*th gene expression intensity x_i. The derivative $\dot{x}_i$ depends on the weighted sum of the expression intensity of all N genes (weights: $a_{i,j}$, total number of genes: N) as well as potentially the weighted sum of external stimuli or inputs u_k (weights: $b_{i,k}$). The ordinary differential equations (ODE) [1] of all genes then form the network model. Possible variants of this description are the usage of difference equations (discrete time model) or non-linear ODEs. If the experiments, the data and thus the model are assumed to be described by steady states, the left hand side of eq. (1) is set to zero:

$$0 = \sum_{j=1}^{N} a_{i,j} x_j + \sum_{k=1}^{M} b_{i,k} u_k \qquad [2]$$

resulting in a static linear model. Sometimes this equation is re-formulated as

$$\hat{x}_i = \sum_{j=1, j\neq i}^{N} a_{i,j} x_j \qquad [3]$$

that describes the steady state expression value of the *i*th gene as the weighted sum (correlation) of all other genes neglecting the explicit formulation of stimulation.

Linear Regression is applied to linear models described by eqs. [1-3]. By means of ordinary least squares (OLS) the solution (estimation $\underline{\hat{\beta}}$ of parameters $\underline{\beta}$) of the model

$$\underline{y} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}$$

with the (normally distributed) error $\underline{\varepsilon}$ can be determined by

$$\underline{\hat{\beta}} = (\underline{\underline{X}}^{\mathrm{T}}\,\underline{\underline{X}})^{-1}\,\underline{\underline{X}}^{\mathrm{T}}\,\underline{y}$$

However, to satisfy the sparseness criterion and other properties of GRNs, methods were developed that mostly comprise the introduction of constraints. Widely used are the ridge regression (Tikhonov regularization), the LASSO (Least Absolute Shrinkage and Selection Operator) regression (Tibshirani, 1996) and LARS (Least-Angle Regression) (Efron et al., 2004). Multiple inference algorithms adapted to the properties of GRNs have been developed, which mostly differ in their ability to infer dynamic models: NIR (Gardner et al., 2003), MNI (di Bernardo et al., 2005), LARS-EN (Zou and Hastie, 2005), TILAR (Hecker et al., 2009b), TIGRESS (Haury et al., 2012) and Kulkarni et al. (2012) (static models) as well as by Gustafsson et al. (2005), TSNI (Bansal et al., 2006), GENLAB (van Someren et al., 2006), Inferelator (Bonneau et al., 2006) and ex-TILAR (Vlaic et al., 2012) (dynamic models).

Complex Optimization

The methods of complex, mostly non-linear optimization extend the linear view of the regression mentioned before. The models are described by non-linear effects, which are the most realistic mechanisms, but many data are needed and high computational effort can be expected. Therefore, some methods have a structure and a parameter optimization step. That class of methods can be further divided into the ones that heuristically try to minimize complexity / computational effort, e.g. NetGenerator (Guthke et al., 2005; Weber et al., 2013) and the ones that pose no assumptions on the model, use sophisticated optimization routines but need many measurement data, e.g. DPLSQ (Nakajima et al., 2012). Further optimization-based methods were described in the works by Mjolsness et al. (2000), Spieth et al. (2006), Nelander et al. (2008) and Hasegawa et al. (2014).

Further Methods

There are further methods that often combine the above mentioned formalisms – motivated by the recent finding that no individual method performs best for all NI tasks (see section *'Assessment of reverse engineer-*

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

ing methods by crowdsourcing'). A method that combines Boolean and probabilistic modeling is PBN (Shmulevich et al., 2002). The algorithm GeneNet (Schäfer et al., 2006) is related to probabilistic modeling but uses partial correlation to infer a Gaussian Graphical Model, a network graph containing both directed and undirected edges. Küffner et al. (2012) developed a method that uses analysis of variance (ANOVA) to generate non-linear correlation networks. A different approach to linear models as the ones mentioned above is the transformation into a model containing the most important components by Singular Value Decomposition (SVD) (Holter et al., 2001; Yeung et al., 2002) and GRN (Wang et al., 2006). The GENIE3 algorithm (Huynh-Thu et al., 2010) infers the GRN by applying tree-based ensemble methods for the selection of predicted interactions for each gene. The algorithm TRaCE performs an ensemble inference of GRNs, which takes into account inherent uncertainty associated with discriminating direct and indirect gene regulations from steady-state data of KO experiments (Ud-Dean and Gunawan, 2014).

Probabilistic modeling and regression was combined in quite different ways by Äijö and Lähdesmäki (2009) as well as Menéndez et al. (2010). While the former developed an algorithm that models the regulatory functions by Gaussian processes, the latter method uses the so-called Graphical LASSO to infer undirected relationships. Yip et al. (2010) presented an approach that uses two different kinds of data (knockout and perturbation), applies probabilistic modeling and optimization on differential equations, respectively and finally combines the results to directed graphs.

The Inferelator introduced by Bonneau et al. (2006), lately was combined with the information theory approach tlCLR to yield dynamic models (Greenfield et al., 2010). It was further combined with other methods to an iterative approach that allows the consideration of prior knowledge and finally generates a consensus networks out of a network ensemble (Greenfield et al., 2013). Mutual information was also combined with an optimization-based approach resulting in the framework NIMOO (Gupta et al., 2011).

While many of the previous methods integrate several results to a final network, some algorithms that could be termed Graph Analysis focus on the elimination of false interactions (pruning of the network). Pinna et al. (2010) used knockout data to establish direct cause-effect relationships and to remove unnecessary feed-forward edges. Flassig et al. (2013) proposed a framework that determines an initial graph from genotype and phenotype correlations and afterwards identifies and removes indirect effects.

GENOME-WIDE VERSUS SMALL- AND MEDIUM-SCALE NETWORKS

Genome-wide networks

Using a holistic approach – in contrast to the reductionist approach – systems biology claims a genome-wide perspective in life sciences exploiting so-called *omics* data that became measurable by high-throughput techniques within the last decades (Sauer et al., 2007). However, genome-wide GRN inference is rarely performed because in general the number of genes and proteins and possible interactions between them in a living organism is much greater than the number of samples and measured data. There are three approaches to tackle this 'curse of dimensionality':

First, the number of measured data can be moderately increased. However, increase of data by interpolation (D'haeseleer et al., 2000) does not introduce additional information from the real biological system which could result in overfitted models. Thus, integration of different data, e.g. from the genome (genotyping, SNPs, TF binding sites, epigenetics), transcriptome, proteome, and interactome (protein-protein, protein-DNA interaction), in particular data from high informative experiments such as external stimulation, knock-out and knock-down experiments, is the best way to tackle the dimensionality problem. On the other hand, generation of more data is of course more ex-

pensive in different extent (Meyer et al., 2014). In addition, integration of heterogeneous data is challenging and requires suitable bioinformatics tools.

Second, knowledge and hypotheses can be introduced to restrict the degree of freedom for GRN modeling. This may be generic or specific. A widely used general assumption is the sparseness of GRNs. This reflects that not all genes and proteins are interconnected. The number of interactions is assumed to be small or only the most important (e.g. strongest) edges are inferred. Next, the number of interactions is reduced in such a manner that the complexity of the inferred network fits to the provided measurements. However, it is known, that there are regulators (e.g. transcription factors, TFs) – often called hubs – that are interlinked with many target genes, i.e. with a high outdegree. Therefore, an enhancement to sparseness is the assumption of scale-freeness (Barabási and Albert, 1999). The outdegree distribution of a scale-free GRN follows a decreasing power law, i.e. the fraction $P(k)$ of nodes having k connections to other nodes is proportional to a power term with the basis k and a negative exponent $-\gamma$ (for a large total number of nodes):

$$P(k) \sim k^{-\gamma}, \quad \gamma > 0$$

Sparseness and scale-freeness are widely used assumptions. Further evolutionary and functional constraints of large- and medium-scale networks are network motifs, robustness, modularity and evolvability (Marbach et al., 2009b). Apart from these generic assumptions, prior knowledge is available and should be exploited from both, databases and literature. Of course, structured information provided in molecular biological databases is preferred over unstructured data in journals and books (see section '*Integration of prior knowledge*').

Currently, both approaches to tackle the problem of dimensionality problem, the extension of measured data and of prior knowledge, may be sufficient for model microorganisms such as the archaeon *Halobacterium* NRC-1 (Bonneau et al., 2006), the prokaryote *Escherichia coli* (Faith et al., 2007; Kaleta et al., 2010) and the eukaryote *Saccharomyces cerevisiae* (Gustafsson and Hörnquist, 2010). For example inferring the genome-wide GRN for *E. coli*, Kaleta et al. discovered the regulation of the lipoate synthase coding gene (lipA) by the pyruvate-sensing pyruvate dehydrogenase repressor (PdhR). First, they used approximately 1.7 million data points from 380 microarray experiments (76 time series with 5 time points each; N = 4,345 genes). Next, as the number of elements $a_{i,j}$ of the full interaction N x N matrix is 18.9 million and exceeds the number of data, they restricted their study to the interaction of the 316 TFs with their potential target gene. Third, the number of the 1.4 million elements of the TF – target gene interaction matrix was reduced to 878 most significant interactions. Among them, 166 edges (19 %) were already known from the RegulonDB (Gama-Castro et al., 2008). Then, using prior knowledge, in this case study focusing on phylogenetic conserved TF binding sites, the percentage of known TF – target gene interactions could be increased from 19 % to 60 % (65/109). Finally, prior knowledge about the metabolic pathway was exploited to select the *in silico* predicted TF – target gene interaction for experimental validation. In total, 23 new targets of the regulator PdhR were discovered by genome-wide NI (Kaleta et al., 2010; Göhler et al., 2011). This large-scale NI was reliable due to the large number of experimental data and the prior knowledge available in databases, including the database RegulonDB as 'gold standard' for assessment of the NI results.

For non-model organisms either experimental data and/or prior knowledge and/or the gold standard are not available in sufficient quantity and/or quality. Thus, genome-wide approaches may lead to GRN of low performance or the performance cannot be assessed. In fact, in most cases the gold standard is simply too small to access performance (as described e.g. for *Staphylococcus aureus* by Marbach et al., 2012). Never-

theless, also in poorly conditioned problems, interesting insights can be gained from medium-scale networks (comprising hundreds of functionally and regulatory characterized genes). Large- and medium-scale networks can also be used to predict potential drug targets and biomarkers for diagnostic purposes and for comparative network analysis (Emmert-Streib et al., 2014 and references therein). For instance, large-scale networks (N > 6,000) for the worm *Caenorhabditis elegans* modeling the correlation between differentially expressed genes were used to study changes of global topological parameters, e.g. the mean node degree under different nutritional conditions during aging (Priebe et al., 2013). For the human pathogenic fungus *Candida albicans* hubs of a 503-gene-network were discussed as known and potential targets of antifungal treatment (Altwasser et al., 2012; Figure 3).

For genome-wide, large-scale modeling, information theory-based methods (e.g. ARACNE) were found to be applicable, however the LASSO-based regression approaches seem to be superior (Altwasser et al., 2012; Meyer et al., 2014).

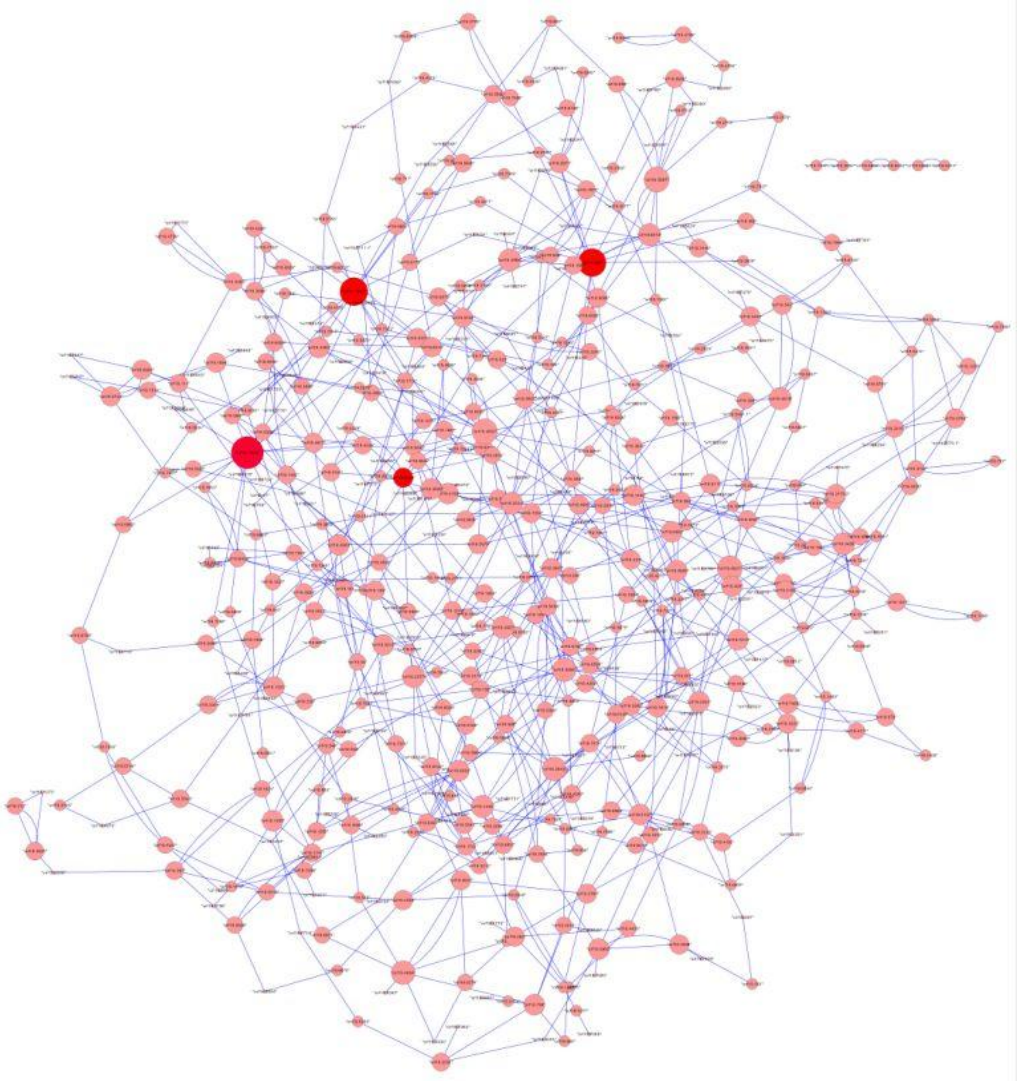

Figure 3: Medium-scale network. 824 interactions inferred using the modified regression method LARS for 503 genes of the 'gold standard' of the human pathogenic fungus *Candida albicans* (Linde et al., 2011, and Altwasser et al., 2012). The red-coloured hubs represent the genes MAL2, SIR2, SNF1 and STE11.

Small-scale networks

In poorly conditioned cases (with respect to the amount of experimental data and prior knowledge), a preferable approach are small-scale networks. The focus is on a subset of genes and proteins and has been demonstrated to be successful for intense interdisciplinary research in biology and medicine. This approach tackles the dimensionality problem by focusing on a subset of genes and proteins, i.e. small-scale modeling instead of the genome-wide approach. The NI of small-scale GRNs is often applied for non-model organisms and tissues. Condensed small-scale GRNs (with up to 50 genes or network nodes) are able to support the experimental design predicting hypotheses of so far unknown mechanisms and interactions in GRNs. Thus, these condensed models could be useful to guide the experimental work (Emmert-Streib et al., 2014).

The main issue of small-scale GRN inference is the feature selection, i.e. the identification of the 'most important' genes or proteins of interest for a certain system or process. For this feature selection there are different approaches. One of them is the clustering of gene expression profiles to select 'representative' nodes (D'haeseleer et al., 2000; Wahde and Hertz, 2000; Mjolsness et al., 2000; Guthke et al., 2005). Alternative or complementary approaches focus on certain functional groups of genes and proteins. The functional groups of interest can be selected by identification of differentially expressed genes (DEGs) followed by gene set enrichment analysis (see section '*Gene enrichment analysis'*). Hypotheses predicted *in silico* using small-scale GRNs were experimentally validated as shown for instance for GRNs describing pathogen-host interaction (Linde et al., 2012; Tierney et al., 2012), adaptation of murine hepatocytes to nutritional change (Vlaic et al., 2012), and mesenchymal stem cell differentiation (Weber et al., 2013). The experimental validation of edges predicted *in silico* by GRN modeling is the best approach for the assessment of NI methods for small-scale GRNs. Here, ODE-based methods such as NetGenerator have been proven to be successfully applicable. Using these tools, hypotheses were predicted *in silico* and validated experimentally afterwards (e.g. Linde et al., 2012; Tierney et al., 2012; Vlaic et al., 2012; Weber et al., 2013). However, these statements about the performance of NI methods may be biased by the specific application. A more objective and generalizable approach has been performed by the so-called DREAM initiative that will be reviewed in the following.

In future, multi-scale modeling by merging modular instead of condensed small- and medium-scale models will open the door to a more holistic approach as claimed in systems biology (Sorger, 2005; Ye et al., 2005).

ASSESSMENT OF REVERSE ENGINEERING METHODS BY CROWDSOURCING

In 2006, Stolovitzky, Monroe and Califano initiated the so-called 'Dialogue for Reverse Engineering Assessment and Methods' (www.dreamchallenges.org) (Stolovitzky et al., 2007; Prill et al., 2010; Marbach et al., 2012; Bansal et al., 2014). From the DREAM initiative an annual research competition is launched (Table 2) and annual DREAM conferences are organized since 2007. Most recently, Califano et al. (2014) reported about the DREAM track of the RECOMB/ISCB Systems and Regulatory Genomics/DREAM Conference 2013.

Community models achieve high performance

Already as the result of the DREAM2 competition, Stolovitzky et al. (2009) resumed that community models, constructed by aggregating predictions across many models submitted by participants achieve performance on a par with the highest-scoring individual models. Remarkably, this high performance is robust to the inclusion of low-scoring models into the ensemble. That finding is important as no individual NI method has been identified showing best performance for all challenges. The superiority

of ensemble learning in GRN inference was confirmed by the following DREAM challenges (Prill et al., 2010, 2011; Marbach et al., 2010; Margolin et al., 2013). Also, summarizing the results of evaluating the GRN inference methods within the DREAM5 competition (Challenge 4 in 2010; Table 2), Marbach et al. (2012) stated that 'the collective knowledge of a community is greater than the knowledge of any individual'. Consequently, methods of combining the information contained within an ensemble of inferred networks were developed (Marbach et al., 2009a).

The goal of the 'DREAM5 – Challenge 4' (Table 2) was to infer an *in silico* benchmark network model (1,643 genes) as well as genome-scale GRN from gene expression microarray datasets for the well-studied microorganisms *E. coli* (4,297 genes, 805 arrays), *S. aureus* (2,677 genes, 160 arrays) and *S. cerevisiae* (5,667 genes, 536 arrays). The evaluation of the *S. aureus* network was excluded from comparison because there are enough experimentally supported interactions ('Gold standard') for network validation available. A total of 35 methods for GRN inference were applied and compared, including regression (8 methods), mutual information (5), correlation (3), Bayesian (6) and other (12) methods as well as combinations of them. The main conclusions from this evaluation of GRN inference methods for genome-scale GRN are the following (Marbach et al., 2012):

(i) After excluding *S. aureus*, the quality of the GRN for *S. cerevisiae* was lowest independent of the applied inference method due to the highest number of genes and relative low number of data.

(ii) Best performance and highest robustness (against poorly performing inference methods) were obtained by community NI approach. In particular, this approach should be preferred for non-model organisms such as *S. aureus* with scarce prior knowledge.

(iii) In general, BN inference methods were outperformed by regression and mutual information approaches (for both *E. coli* and *in silico* GRN). It is known that BN need more data than other NI methods. Integration of prior knowledge, which is an important advantage of BN, was not requested in the 'DREAM5 – Challenge 4'.

(iv) Certain regression methods show a performance similar to the best-performing community NI. The well-established mutual information NI methods CLR and ARACNE are outperformed by certain LASSO/LARS-based regression methods. The method TIGRESS (Haury et al., 2012) combined LARS with a novel feature selection method ('stability selection'). However, LASSO combined with bootstrapping, which was found to be the best performing individual method for the *in silico* NI, achieved only a low score for the *E. coli* GRN.

(v) A specific inference method (Küffner et al., 2012) outperformes the community NI – but only for the *E. coli* GRN inference: Here, co-dependencies between TFs and target genes are detected by two-way ANOVA method. In addition, TF perturbation data are up-weighted. For the *E. coli* NI also very good results were obtained using the LASSO toolbox GENLAB with default parameters (http://genlab.tudelft.nl/genlab.html; van Someren et al., 2006).

(vi) The random forest-based method GENIE3 was the best performer in the 'DREAM4 – Challenge 2' for *in silico* NI inference (Huynh-Thu et al., 2010) and reached also a high score in the DREAM5 – 4 challenge for both *in silico* and *E. coli* NI. The random forest-based method GENIE3 was ranked with highest overall score in the evaluation of the 'DREAM 5 - Challenge 4' (Marbach et al., 2012). Here, decision trees are used to produce prioritized lists of TFs regulating each target gene.

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

Table 2: DREAM Challenges. DREAM#: Running number of challenge; Short Title and reference if any; Data given for the challenge

Year DREAM# - Challenge#	Short Title (Reference)	Data	Task/ Goal
2006 1	Stolowitzky et al., 2007		
2007 2	Stolowitzky et al., 2009		
2007 2 - 1	Transcriptional target prediction	ChIP-on-chip data of 200 genes after perturbation of BCL2-pathway in B-cells	Predict the genes for TF binding
2007 2 – 2	Protein-Protein interaction network	Y2H data	Predict a PPI network of 47 proteins
2007 2 – 3	Synthetic five-gene network inference Cantone et al., 2009 Marbach et al., 2009a, b Äijö and Lähdesmäki, 2009	QPCR and gene expression time series after 2 treatments (588*10*2) Chip data of *in vivo* model organism	Infer a gene regulation network from qPCR and microarray measurements
2007 2 – 4	*In silico* network Gustafsson et al., 2009	Simulated time series from three *in silico* 50-gene-GRNs (50*26*23)	Infer various network topologies and connectivity of the three GRNs
2007 2 - 5	Gene-scale network	Microarray data from a microorganism	Reconstruct a genome scale regulatory network from a large collection of microarrays
2008 3	Prill et al., 2010		
2008 3 - 1	Signaling cascade identification	Incomplete flow cytometry data	Infer a signaling network
2008 3 – 2	Signaling response prediction Prill et al., 2011 Guex et al., 2010 Clarke et al., 2010	Phosphoproteomics data	Predict missing protein concentrations from a large corpus of measurements
2008 3 – 3	Gene expression prediction Gustafsson and Hörnquist, 2010 Ruan, 2010	Gene expression time course data for four different strains of yeast (*S. cerevisiae*), after perturbation of the cells	Predict missing gene expression measurements

Table 2 (cont.): DREAM Challenges. DREAM#: Running number of challenge; Short Title and reference if any; Data given for the challenge

Year DREAM# - Challenge#	Short Title (Reference)	Data	Task/ Goal
2008 3 - 4	*In silico* network Marbach et al., 2010 Yip et al., 2010 Madar et al., 2010	10, 50, 100 gene time series with 21 time points	Infer simulated gene regulation networks
2009 4 – 1	Peptide Recognition Domain Specificity prediction	5 human SH3 domain sequences, 3 serine/threonine kinase sequences and 5 synthetic PDZ domain sequences modeled on Erbin (Erbb2 interacting protein)	Predict protein-protein interactions at the level of binding domains and peptides
2009 4 – 2	*In silico* network challenge Huynh-Thu et al., 2010 Menéndez et al., 2010 Pinna et al., 2010	Simulated steady-state and 10 time-series data (10 and 100 genes; 21 time points; wild-type, knockouts, knockdowns, multifactorial perturbations)	Infer *in silico* GRN and predict gene expression measurements in response to perturbations
2009 4 – 3	Predictive signaling network modeling Eduati et al., 2010 Prill et al., 2011	Activity levels of signaling proteins in HepG2 cell lines	Predict phosphoprotein measurements using an interpretable, predictive network
2010 5	Marbach et al., 2012		
2010 5 – 1	Epitope-Antibody Recognition	Sequences of peptides that either bind intravenous immunoglobulin antibodies or do not	Predict the binding specificity of peptide antibody interactions
2010 5 – 2	TF-DNA Motif Recognition Weirauch et al., 2013	Protein Binding Microarray data (41,000 60-base probe sequences)	Predict the specificity of a TF-binding to a 35-mer probe
2010 5 – 3	Systems Genetics Loh et al., 2011 Vignes et al., 2011	*In silico* (1000 gene-networks) and experimental (soybean) genotype and gene expression data	Predict disease phenotypes and infer gene networks

Table 2 (cont.): DREAM Challenges. DREAM#: Running number of challenge; Short Title and reference if any; Data given for the challenge

Year DREAM# - Challenge#	Short Title (Reference)	Data	Task/ Goal
2010 5 - 4	Network inference Marbach et al., 2012 Haury et al., 2012	Four microarray data sets (hundreds experiments each), three from pathogenic microorganisms after perturbation with drugs etc, one from *in silico* network	Infer simulated and *in vivo* GRNs
2011 6 – 1	Alternative Splicing	Short-read mRNA-Seq data	Reconstruct Alternative Splicing mRNA transcripts
2011 6 – 2	Estimation of Model Parameters Meyer et al., 2014	Three GRNs	Iterative optimization of kinetic parameter values and experimental design
2011 6 – 3	Gene Expression Prediction	Promoter sequences in eukaryotes	Predict gene expression levels
2011 6 – 4	Molecular classification of AML	Patient samples using flow cytometry data	Diagnose AML
2012 7 –1	Network Topology & Parameter Inference Meyer et al., 2014	Structure of 9-gene GRN, incomplete 11-gene GRN	Parameter values; missing links; predict outcomes of perturbations
2012 7 –2	Breast cancer prognosis Margolin et al., 2013	Clinical information about the tumor and genome-wide molecular profiling data including gene expression and copy number profiles of 1981 patients	Assess the accuracy of computational models designed to predict breast cancer survival
2012 7 –3	ALS prediction Küffner et al., 2015	1,822 ALS patient's disease status during 3 months after diagnosis: demographics, medical and family history data, functional measures, vital signs, and lab data	Predict the future progression of disease in ALS patients 12 months after the diagnosis
2012 7 –4	Drug Sensitivity Prediction Costello et al., 2014	Genomic, epigenomic and proteomic profiling data sets measured in human breast cancer cell lines	Predict drug sensitivity and synergies in breast cancer cell lines

Table 2 (cont.): DREAM Challenges. DREAM#: Running number of challenge; Short Title and reference if any; Data given for the challenge

Year DREAM# - Challenge#	Short Title (Reference)	Data	Task/ Goal
2013 8 – 1	Breast cancer Network Inference	45 proteomics and 125 protein time-course datasets on four breast cancer cell lines	NI and prediction of timecourse
2013 8 – 2	Toxicogenetics	Genetics and transcriptomics information of the 1000 Genomes Project; cytotoxicity measures for > 100 toxic agents	Model cytotoxicity across cell lines; predict absolute cytotoxicity for which cytotoxicity data are not provided; predict median, 5%-quantile, and 95%-quantile EC10 across the population for each of 50 unknown compounds
2013 8 – 3	Whole-cell parameter estimation	Whole cell model of *Mycoplasma genitalium*; simulated data	Estimating the model parameters for specific biological processes from simulated data
2014 9 - 1	Essentiality Prediction	Gene expression and/or gene copy number features	predictive models to infer genes that are essential to cancer cell viability
2014 9 - 2	AML Outcome prediction	Clinical cytogenetics, known genetics markers and phosphoproteomic data	Predict the outcome of treatment of AML patients
2014 9 - 3	Alzheimer's Disease Big Data	Clinical data of a longitudinal multicenter study, >1600 participants	Predict the best biomarkers for early AD-related cognitive decline and for the mismatch between high amyloid levels and cognitive decline
2014 9 - 4	Somatic mutation calling	Whole-genome sequencing data from tumor and normal samples	Predict mutation calls associated with cancer; identify the most accurate mutation detection algorithms
2014 9 - 5	Rheumatoid Arthritis Responder Plenge et al., 2013	Clinical, genotyping (by Affymetrix chips) and drug dosage data from 2,706 individuals	Predict anti-TNF response in Rheumatoid Arthritis

Integration of heterogeneous data

Both, the 'DREAM7 – Challenge 1' and the 'DREAM6 – Challenge 2', aim to evaluate methods for model structure discrimination and for the estimation of parameters in non-linear biochemical models that characterize the dynamics of molecular processes (Meyer et al., 2014). In the 'DREAM7 – Challenge 1', an 9-gene network composed *in silico* was used as gold standard for pa-

rameter estimation. Additionally, an incomplete 11-gene network model was used as gold standard to assess methods for identification of three missing links. A virtual budget was provided in this challenge to 'buy' experimental data generated (*in silico*) by model simulations. The expense of different experimental techniques, such as transcriptome profiling measured by microarrays as well as abundance of all proteins measured by mass spectrometry, both with low and high temporal resolution (500 and 100 credits, respectively) and protein abundance for 2 proteins of choice with highest temporal resolution (for 400 credits) was mimicked. Furthermore, perturbation experiments, such as knock-out, knock-down and decrease of ribosome binding site experiments are offered for choice (for 800, 350, and 450 credits, respectively). The analysis of the results from 19 competing teams suggests that the combination of state-of-the-art parameter estimation and a varied set of experimental methods using a few datasets, mostly proteome (fluorescence imaging) data, can accurately determine parameters of biochemical models of gene regulation.

The identification of the missing links in the incomplete 11-gene network was more challenging. For identifying the missing links of the 11-gene network, the best-performing team first used credits on wild type fluorescence data, to cheaply obtain a setting with qualitative disagreement between data and model, and then used mass spectroscopy experiments with perturbations to test for potential missing links. For discrimination between the alternative model structures, Meyer et al. (2014) applied classical maximum likelihood methods. However, only the consensus obtained by majority voting to select the most submitted links had a top performing score. Only one of the three consensus links was correctly inferred, while the direction and nature of the regulatory link of the two others were incorrect. This demonstrates the difficulty to correctly identify the topology also of small-scale GRNs based solely on limited experimental data and perturbations. Thus, integration of prior knowledge is indispensable also for small-scale GRN inference. Again, it was found that aggregating independent parameter predictions and network topologies across submissions created a solution that can be better than the one from the best-performing submission (Meyer et al., 2014).

The DREAM challenges were not only focused on the assessment of GRN inference methods. As the consequence of the conclusion that genome-wide GRN modeling is reliable only for well-conditioned problems and that the identification of the GRN topology may be wrong also for small-scale GRN (if no prior knowledge was included), more simplified but useful computational problems were addressed by the DREAM competition. Thus, there were also challenges to evaluate methods for gene expression and biomarker prediction as well as to assess the performance of classifiers for diagnosis of diseases. Ruan (2010) showed for the prediction of gene expression values that the simple k-nearest-neighbor method led to almost the same performance as a much more sophisticated method.

In the following, three examples from DREAM7 will be discussed that provide recommendations for inference of predictive models in a clinical perspective.

In the 'DREAM7 – Challenge 2' (Table 2), the results of breast cancer survival prediction by more than 1,400 computational models from 354 research groups were evaluated. The models were trained on the data set of 1,000 samples including clinical information (for example, age, tumor size and histological grade), mRNA expression data and DNA copy number data. The models were validated on data sets of 981 samples. The predictive value of each model was scored by calculating the concordance index (CI) of predicted death risk. In a final phase, the data of all 1,981 samples were used for model refinement and the retrained models were validated on a further data set from 184 women diagnosed with breast cancer. The best-performing model combined clinical

features and molecular features selected by prior knowledge. A machine learning method (boosted regression) was applied to a combination of clinical features, expression levels of genes selected by data-driven criteria and by their involvement in breast cancer (Ravasi et al., 2010) and, finally, an aggregated "genomic instability" index calculated from the copy number data (Bilal et al., 2013). Margolin et al. (2013) found that the top-scoring models used a methodology that minimized overfitting to the training set by defining a "Metagene" feature space based on robust gene expression patterns observed in multiple external cancer data sets. Long-time survivors are better predicted than short-time survivors.

The goal of the 'DREAM7 – Challenge 3' was to predict the future progression of disease in Amyotrophic Lateral Sclerosis (ALS), a neurodegenerative disease. The results of 37 algorithms were submitted for evaluation. Interestingly, the two best algorithms outperformed predictions by ALS clinicians. Küffner et al. (2015) estimated that using both winning algorithms in future trial designs could reduce the required number of patients by at least 20 %. In addition, several potential non-standard predictors of ALS progression were identified including uric acid, creatinine and blood pressure. Thus, this DREAM challenge contributed to a better understanding of ALS pathobiology.

The 'DREAM7 – Challenge 4' dealt with prediction of the drug response from multi-omics data (CNV, RPPA, Methylation, Exome sequencing, gene expression microarray, RNA-seq) measured in human breast cancer cell lines, a total of 44 drug sensitivity prediction algorithms were analyzed (Costello et al., 2014). The best results were obtained using nonlinear relationships and incorporating the biological pathway information. In addition, they found that gene expression microarrays provided the best predictive power, however, the performance increased including further data sets. The top-3 approaches used Bayesian multitask Multiple Kernel Learning, weighted nonlinear regression trees and weighted features from Pearson's correlation.

CHALLENGES

Integration of prior knowledge

The prediction of GRNs is a great combinatorial challenge usually based on a limited amount of data. Some inference tools integrate prior knowledge to support the inference process. Supportive interaction knowledge is usually of positive nature, meaning that the existence of an interaction was predicted or experimentally observed. Negative prior knowledge about non existing interactions is equally important for network inference, but hardly available. One reason is that the experimental proof that two genes never interact is very hard to do. Some journals publish negative results (Journal of Negative Results in Biomedicine; New Negatives in Plant Science (Elsevier); Journal of Negative Results – Ecology & Evolutionary Biology) from which negative prior knowledge can be extracted. Nevertheless, to our knowledge no databases exist that allow easy access to negative prior knowledge.

There is an increasing number of molecular biological databases. Currently, the Nuclear Acid Research Online Molecular Biology Database Collection has been expanded to 1,552 databases (Fernández-Suárez et al., 2014). Olsen et al. (2014) assessed the relevance of different prior knowledge sources for inferring GRNs in cancer research. The most direct interactions that can be obtained as prior knowledge for GRNs are TFs interacting with promoters. Promoters can be analyzed for known TF binding profiles (free JASPAR database (Mathelier et al., 2013), commercial TRANSFAC database (Matys et al., 2006)) or motifs can be elicited (e.g. MEME (Bailey et al., 2009)). Experimentally, TF-DNA interactions are determined by ChIP-Seq resulting in p-values of interactions. These p-values are inversely correlated to the probability of an edge being present in a GRN (Bernhard and Hartemink, 2005).

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

Various databases have been established to provide knowledge about genes and interactions. Some include many organisms (e.g. STRING (Szklarczyk et al., 2015)), others are species specific such as the Candida Genome Database (Binkley et al., 2014). Over the last years, pathogen-host interaction databases emerged and are listed in Mukherjee et al. (2013). One example is PHISTO (Tekir et al., 2013), which includes >16,000 virus-host interactions, > 8,000 bacteria-host interactions but only few fungus-host interactions (<10). Especially for not well studied organisms available knowledge can be limited and insufficient. One option would be to extract knowledge of homologous genes in closely related organisms.

Human database curation teams cannot keep up with volume and pace of literature production (Baumgartner et al., 2007). Thus, text mining needs to close this gap by extracting structured knowledge from unstructured information such as scientific literature (Hahn et al., 2007). Text-mining tools as for instance the commercial Pathway Studio for mammals (Nikitin et al., 2003) or the free Gene Interaction Miner (Ikin et al., 2010) facilitate automatic knowledge extraction from literature databases. For well studied organisms this can result in a lot of interactions, for which manual curation might not be feasible anymore.

Furthermore, more knowledge is available for well studied genes which has to be taken into account when predicted GRNs are interpreted. Also, web applications exist that retrieve knowledge from various sources (Haibe-Kains et al., 2012; Horn et al., 2014). Given a set of genes, the web platform GeneMANIA (Mostafavi et al., 2008; Montojo et al., 2014) queries biological databases, published articles and co-expression from published data sets. It returns a network of interactions distinguishable by source and is available for *H. sapiens* and eight model organisms (e.g. *Mus musculus*, *S. cerevisiae*).

Pathogen-host GRNs

Next Generation Sequencing techniques paved the way for advanced genomic and transcriptomic studies of interacting species, in particular for metagenomic and infection research (Pallen et al., 2010). Dual RNA-Seq is an approach, where transcriptomes of two or more species are sequenced together. Westermann et al. (2012) reviewed various aspects that have to be considered when dual RNA-Seq samples are prepared and sequenced.

One application of dual RNA-Seq is to study pathogens interacting with their host, e.g. filarial worm – mosquito interactions (Choi et al., 2014) and *Azospirillum brasilense* colonizing wheat roots (Camilios-Neto et al., 2014). Furthermore, Tierney et al. (2012) published a dual RNA-Seq time series data set of mouse dendritic cells infected with *C. albicans*. Six pathogen genes and five host genes were selected and the small-scale NI tool NetGenerator was applied. Two of the predicted interactions were experimentally validated demonstrating the applicability of GRN inference to model pathogen-host interactions. Recently, NetGenerator was extended and its application to infer pathogen-host GRNs was outlined (Schulze et al., 2015). The focus was on accounting for pathogen-host interaction data characteristics, such as changing environmental conditions, temporally different onsets of transcriptional responses and possible missing data points (e.g. only one organism survives). Furthermore, they give an overview of basic requirements and main steps of acquisition and analysis of dual RNA-Seq data.

Non-linear models

Based on fundamental knowledge from thermodynamics of irreversible processes and self-organization (Prigogine and Nicolis, 1971) it is known that living systems have to be modeled using non-linear functions. However, adequate non-linear modeling requires more experimental data or/and prior knowledge, which is in most cases already

EXCLI Journal 2015;14:346-378 – ISSN 1611-2156
Received: January 29, 2015, accepted: February 10, 2015, published: March 02, 2015

insufficient for linear modeling (problem of dimensionality). The DREAM2 challenge provided an *in silico* network with 50 genes and a 'comfortable' set of 50 x 26 x 23 data without noise and the result showed that also the best-performing LASSO-based NI with a set of non-linear basis functions does not correctly infer the non-linear relations (Gustafsson et al., 2009). To include non-linearity in the GRN but apply the benefits of linear modeling, piecewise linear models were proposed (Westra et al., 2011), but they are not extensively studied for GRN inference so far. Thus, linear models are often applied despite the knowledge that the assumption of linearity is not 'true'. Linear models may be wrong, but useful for prediction of hypotheses (e.g. regulator – target gene relations) or the behavior of the system close to the steady state (e.g. outcome of a disease). However, in general, such modeling is insufficient to predict a molecular mechanism in detail and for simulation of the dynamic behavior with multiple attractors (e.g. multiple steady states; Milnor, 1985). Small- and medium-scale models are helpful to support and design experiments. To overcome the reductionistic view and go ahead to a more holistic one, multi-scale modeling merging various validated small-scale models will be the future in systems biology of gene expression network modeling.

Model validation

Most important, GRN results have to be validated experimentally. Typically, this is an iterative process in systems biology forming a cycle of wet-lab and dry-lab research (Figure 1). Based on experimental data and prior knowledge at hand, an initial network model can be inferred. Then, hypotheses have to be generated based on the draft model and should be experimentally tested or checked by literature search. In general, this experimental validation or deep literature search will give rise to refined network modeling. For this second round of network modeling, the results of experimental validation or literature search can be included as improved prior knowledge. Of course, some prior knowledge relations can be contradictory to each other or to the experimental data. Contradictory prior knowledge can be handled by ranking it with a score according to the experimental method and setup used for drawing the respective conclusion. The score can also reflect how close the experimental setup of the referenced system is to the currently studied system (with respect to experimental conditions, tissues, organisms). Such scoring systems are in its infancy (e.g. Linde et al., 2010).

CONCLUSION

The mathematical and computational modeling of networks is of great importance in biomedical research to understand molecular mechanisms, e.g. pathogen-host interaction. To tackle the complexity inherent in large networks of interacting biomolecules, many different approaches and methods have been established. They include methods based on Boolean networks, Bayesian Networks, information theory, differential or difference equations, graphical Gaussian models and/or supervised machine learning methods. In general due to the high dimensionality (thousands of genes and proteins) versus the limited number of samples (not more than hundreds), the GRN inference is underdetermined implying that there could be many equivalent (indistinguishable) solutions. To cope with this fundamental problem, there are various approaches for GRN inference. Some of them are widely used and powerful, such as the information theory-based methods (like the ARACNE or Context Likelihood of Relatedness – CLR method) and the regression-based LASSO method for large-scale network models, whereas ODE-based methods, such as Inferelator and NetGenerator, for dynamical medium- and small-scale models, respectively. During the last decade the scientific community improved the understanding when and how to apply them. As a trend of the last five years, different methods of NI and different data types, including prior knowledge, were com-

bined and integrated to improve the performance.

First of all, no individual GRN modeling approach performs best for all problems. The DREAM competition showed that ensemble learning, i.e. 'community models' constructed by aggregating predictions across many models, allows NI with high performance and robustness against the inclusion of low-performance models. Integration of predictions from multiple inference methods is robust and has high performance across diverse datasets (Marbach et al., 2009a; 2012). For instance, the algorithm TRaCE performs an ensemble inference of GRNs (Ud-Dean and Gunawan, 2014).

Second, NI should not be based only on the limited number of gene expression data. GRN inference should integrate prior knowledge (see section '*Integration of prior knowledge*') or further, heterogeneous experimental data sets (Greenfield et al., 2010). In particular, GRN inference using gene expression data should be supported by information about already known TF – target gene interaction, TF binding sites or their motifs, composites of TFs and signaling pathways from receptors to TFs. Marbach et al. (2009b) won the synthetic five-gene network challenge of the reverse engineering competition inference DREAM2 by integration of prior knowledge mimicking the evolutionary process.

Third, to infer networks with directed edges, data should be exploited that represent cause–effect relations. Typically, these are time series data of response to known perturbations (interventions), steady state data from knock-out (KO) experiments or assigned with Singular Nucleotide Polymorphism (SNP) or other genotype data. The type of these data is essential for the choice of inference method. For small-scale network modeling using time series data and prior knowledge, ODE-based inference tools, such as NetGenerator (Weber et al., 2013), should be preferentially applied. For medium-scale models based on time series and mixed data including steady state data, the ExTILAR algorithm (Vlaic et al., 2012) was developed. For large-scale network modeling using time series data, TimeDelay-ARACNE is applicable (Zoppoli et al., 2010). However, for genome-wide and other large-scale networks the regression-based method LASSO (GENLAB, van Someren et al., 2006) seems to be best situated if it is well configured and the experimental data and prior knowledge are of sufficient quantity and quality (Marbach et al., 2012). Boolean and static network modeling should be preferred if the data are mainly steady state gene expression data from KO experiments or for modeling of signaling pathways, respectively (Eduati et al., 2010; Klamt et al., 2010; Flassig et al., 2013; Samaga and Klamt, 2013; Pinna et al., 2013; Ryll et al., 2014; Nakajima and Akutsu, 2014).

Today, large-scale models for model organisms (e.g. *E. coli* and *S. cerevisiae*) reflect some general properties such as the robustness and stability of the system or they give information about hubs, e.g. most important transcriptional regulators and target genes. However, they do not comprehensively represent all interactions and their dynamics. Currently, for non-model organisms, small-scale networks with a certain focus, is a useful approach. However, we expect that in future also for non-model organisms genome-wide GRN models with improved predictive power will be established based on extended experimental data compendia and molecular interactome databases.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 'Pathogenic fungi and their human host: Networks of interaction', subproject INF (JL, RG) and B3 (SS, RG). SGH was supported by the German Federal Ministry of Education and Research (BMBF) within the Virtual Liver initiative.

REFERENCES

Äijö T, Lähdesmäki H. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. Bioinformatics. 2009; 25:2937-44.

Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput. 1999:17-28.

Altay G, Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks. BMC Syst Biol. 2010;4:132.

Altwasser R, Linde J, Buyko E, Hahn U, Guthke R. Genome-wide scale-free network inference for *Candida albicans*. Front Microbiol. 2012;3:51.

Ay A, Arnosti DN. Mathematical modeling of gene expression: a guide for the perplexed biologist. Crit Rev Biochem Mol Biol. 2011;46:137-51.

Bailey TL, Bodén M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37(Suppl 2):W202-8.

Bansal M, di Bernardo D. Inference of gene networks from temporal gene expression profiles. IET Syst Biol. 2007;1:306-12.

Bansal M, Della Gatta G, di Bernardo D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics. 2006;22:815-22.

Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007;3:78. Erratum in: Mol Syst Biol. 2007;3:122.

Bansal M, Yang J, Karan C, Menden MP, Costello JC, Tang H, et al; NCI-DREAM Community. A community computational challenge to predict the activity of pairs of compounds. Nat Biotechnol. 2014;32:1213-22.

Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509-12.

Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101-13.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005;37:382-90.

Baumgartner WA Jr, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics. 2007;23:i41-8.

Bernhard A, Hartemink AJ. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. Pac Symp Biocomput. 2005;2005:459-70.

Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, et al. Improving breast cancer survival analysis through competition-based multidimensional modeling. PLoS Comput Biol. 2013;9(5):e1003047.

Binkley J, Arnaud MB, Inglis DO, Skrzypek MS, Shah P, Wymore F, et al. The Candida Genome Database: the new homology information page highlights protein similarity and phylogeny. Nucleic Acids Res. 2014;42(Database issue):D711-6.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-21.

Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. Genome Biol. 2006;7(5):R36.

Butcher EC, Berg EL, Kunkel EJ. Systems biology in drug discovery. Nat Biotechnol. 2004;22:1253-9.

Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. 2000:418-29.

Califano A, Kellis M, Stolovitzky G. RECOMB/ISCB systems biology, regulatory genomics, and DREAM. 2013 special issue. J Comput Biol. 2014;21:371-2.

Camilios-Neto D, Bonato P, Wassem R, Tadra-Sfeir MZ, Brusamarello-Santos LCC, Valdameri G, et al. Dual RNA-seq transcriptional analysis of wheat roots colonized by *Azospirillum brasilense* reveals up-regulation of nutrient acquisition and cell cycle genes. BMC Genomics. 2014;15:378.

Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, et al. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. Cell. 2009;137:172-81.

Choi YJ, Aliota MT, Mayhew GF, Erickson SM, Christensen BM. Dual RNA-seq of parasite and host reveals gene expression dynamics during filarial worm–mosquito interactions. PloS Negl Trop Dis. 2014;8(5):e2905.

Clarke ND, Bourque G. Success in the DREAM3 signaling response challenge using simple weighted-average imputation: lessons for community-wide experiments in systems biology. PLoS One. 2010;5(1): e8417.

Compston A, Coles A. Multiple sclerosis. Lancet. 2008;372(9648):1502-17.

Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al.; NCI DREAM Community. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32:1202-12.

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. 2013;8(12): e85024.

de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol. 2002;9:67-103.

D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics. 2000;16:707-26.

di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, et al. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat Biotechnol. 2005;23:377-83.

Eduati F, Corradin A, Di Camillo B, Toffolo G. A Boolean approach to linear prediction for signaling network modeling. PLoS One. 2010;5(9):e12789.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Statist. 2004;32:409–99.

Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Front Cell Dev Biol. 2014;2:38.

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al.; RGASP Consortium. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013;10:1185-91.

Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z. Reconstructing dynamic regulatory maps. Mol Syst Biol. 2007;3:74.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):e8.

Fernández-Suárez XM, Rigden DJ, Galperin MY. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. Nucleic Acids Res. 2014;42(Database issue):D1-6.

Finotello F, Lavezzo E, Bianco L, Barzon L, Mazzon P, Fontana P, et al. Reducing bias in RNA sequencing data: a novel approach to compute counts. BMC Bioinformatics. 2014;15(S-1):7.

Flassig RJ, Heise S, Sundmacher K, Klamt S. An effective framework for reconstructing gene regulatory networks from genetical genomics data. Bioinformatics. 2013;29:246-54.

Follo MY, Manzoli L, Poli A, McCubrey JA, Cocco L. PLC and PI3K/Akt/mTOR signalling in disease and cancer. Adv Biol Regul. 2015;57:10-6.

Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000;7:601-20.

Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spinola MI, Contreras-Moreira B, et al. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res. 2008;36(Database issue):D120-4.

Gardner TS, Faith JJ. Reverse-engineering transcription control networks. Phys Life Rev. 2005;2:65-88.

Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003;301(5629):102-5.

Glocker MO, Guthke R, Kekow J, Thiesen HJ. Rheumatoid arthritis, a complex multifactorial disease: on the way toward individualized medicine. Med Res Rev. 2006;26:63-87.

Göhler AK, Kökpinar Ö, Schmidt-Heck W, Geffers R, Guthke R, Rinas U, et al. More than just a metabolic regulator - elucidation and validation of new targets of PdhR in Escherichia coli. BMC Syst Biol. 2011;5: 197.

Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. BMC Syst Biol. 2014;8 (Suppl 2):I1.

Greenfield A, Madar A, Ostrer H, Bonneau R. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. PLoS One. 2010;5(10):e13397.

Greenfield A, Hafemeister C, Bonneau R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics. 2013;29:1060-7.

Groisman EA, Mouslim C. Sensing by bacterial regulatory systems in host and non-host environments. Nat Rev Microbiol. 2006;4:705-9.

Guex N, Migliavacca E, Xenarios I. Multiple imputations applied to the DREAM3 phosphoproteomics challenge: a winning strategy. PLoS One. 2010;5(1): e8012.

Gupta R, Stincone A, Antczak P, Durant S, Bicknell R, Bikfalvi A, et al. A computational framework for gene regulatory network inference that combines multiple methods and datasets. BMC Syst Biol. 2011;5: 52.

Gustafsson M, Hörnquist M. Gene expression prediction by soft integration and the elastic net-best performance of the DREAM3 gene expression challenge. PLoS One. 2010;5(2):e9134.

Gustafsson M, Hörnquist M, Lombardi A. Constructing and analyzing a large-scale gene-to-gene regulatory network--lasso-constrained inference and biological validation. IEEE/ACM Trans Comput Biol Bioinform. 2005;2:254-61.

Gustafsson M, Hörnquist M, Lundström J, Björkegren J, Tegnér J. Reverse engineering of gene networks with LASSO and nonlinear basis functions. Ann N Y Acad Sci. 2009;1158:265-75.

Guthke R, Möller U, Hoffmann M, Thies F, Töpfer S. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. Bioinformatics. 2005;21:1626-34.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494-512.

Hahn U, Wermter J, Blasczyk R, Horn PA. Text mining: powering the database revolution. Nature. 2007;448(7150):130.

Haibe-Kains B, Olsen C, Djebbari A, Bontempi G, Correll M, Bouton C, et al. Predictive networks: a flexible, open source, web application for integration and analysis of human gene networks. Nucleic Acids Res. 2012;40(Database issue): D866-D875

Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac Symp Biocomput. 2001:422-33.

Hasegawa T, Yamaguchi R, Nagasaki M, Miyano S, Imoto S. Inference of gene regulatory networks incorporating multi-source biological knowledge via a state space model with L1 regularization. PLoS One. 2014;9(8):e105942.

Haury AC, Mordelet F, Vera-Licona P, Vert JP. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. BMC Syst Biol. 2012;6:145.

Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models - a review. Biosystems. 2009a;96:86-103.

Hecker M, Goertsches RH, Engelmann R, Thiesen HJ, Guthke R. Integrative modeling of transcriptional regulation in response to antirheumatic therapy. BMC Bioinformatics. 2009b;10:262.

Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR. Dynamic modeling of gene expression data. Proc Natl Acad Sci U S A. 2001;98:1693-8.

Horn F, Rittweger M, Taubert J, Lysenko A, Rawlings C, Guthke R. Interactive exploration of integrated biological datasets using context-sensitive workflows. Front Genet. 2014;5:21.

Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1-13.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;5(9): e12776.

Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet. 2001;2:343-72.

Ikin A, Riveros C, Moscato P, Mendes A. The Gene Interaction Miner: a new tool for data mining contextual information for protein-protein interaction analysis. Bioinformatics. 2010;26:283-4.

Isci S, Dogan H, Ozturk C, Otu HH. Bayesian network prior: network analysis of biological data using external knowledge. Bioinformatics. 2014;30:860-7.

Ivashchenko AT, Tauasarova MI, Atambayeva ShA. Exon-intron structure of genes in complete fungal genomes. Molecular Biol. 2009;43:24-31.

Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014; 15(1):18.

Kaleta C, Göhler A, Schuster S, Jahreis K, Guthke R, Nikolajewa S. Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis. BMC Syst Biol. 2010;4:116.

Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol. 1969;22:437-67.

Klamt S, Flassig RJ, Sundmacher K. TRANSWESD: inferring cellular networks with transitive reduction. Bioinformatics. 2010;26:2160-8.

Koshland DE Jr. The seven pillars of life. Science. 2002;295(5563):2215-6.

Küffner R, Petri T, Tavakkolkhah P, Windhager L, Zimmer R. Inferring gene regulatory networks by ANOVA. Bioinformatics. 2012;28:1376–82.

Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nat Biotechnol. 2015;33:51-7.

Kulkarni VV, Arastoo R, Bhat A, Subramanian K, Kothare MV, Riedel MC. Gene regulatory network modeling using literature curated and high throughput data. Syst Synth Biol. 2012;6(3-4):69-77.

Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, et al. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 2014;15(6):R86.

Leclerc RD. Survival of the sparsest: robust gene networks are parsimonious. Mol Syst Biol. 2008;4: 213.

Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, et al. The systems biology graphical notation. Nat Biotechnol. 2009;27:735-41. Erratum in: Nat Biotechnol. 2009;27:864.

Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput. 1998:18-29.

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30:923-30.

Linde J, Wilson D, Hube B, Guthke R. Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells. BMC Syst Biol. 2010;4:148.

Linde J, Buyko E, Robert A, Hahn U, Guthke R. Full-genomic network inference for non-model organism: A case study for the fungal pathogen Candida albicans. In: International Conference on Systems Biology (ICSB-2011). Heidelberg – Mannheim, Germany, August 28th – September 1st, 2011.

Linde J, Hortschansky P, Fazius E, Brakhage AA, Guthke R, Haas H. Regulatory interactions for iron homeostasis in Aspergillus fumigatus inferred by a Systems Biology approach. BMC Syst Biol. 2012;6:6.

Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2013;30:301-4.

Lo K, Raftery AE, Dombek KM, Zhu J, Schadt EE, Bumgarner RE, et al. Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. BMC Syst Biol. 2012;6:101.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J: Molecular cell biology, 4th ed (section 11.6). New York: W.H. Freeman, 2000.

Loh PR, Tucker G, Berger B. Phenotype prediction using regularized regression on genetic data in the DREAM5 Systems Genetics B Challenge. PLoS One. 2011;6(12):e29095.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R. DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. PLoS One. 2010;5(3):e9803.

Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biology. 2011;9(1):34.

Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-seq. Med Sci Monit Basic Res. 2014;20:138–41.

Marbach D, Mattiussi C, Floreano D. Combining multiple results of a reverse-engineering algorithm: application to the DREAM five-gene network challenge. Ann N Y Acad Sci. 2009a;1158:102-13.

Marbach D, Mattiussi C, Floreano D. Replaying the evolutionary tape: biomimetic reverse engineering of gene networks. Ann N Y Acad Sci. 2009b;1158:234-45.

Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci U S A. 2010;107:6286-91.

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9:796-804.

Margolin AA, Bilal E, Huang E, Norman TC, Ottestad L, Mecham BH, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Sci Transl Med. 2013;5(181):181re1.

Markowetz F, Bloch J, Spang R. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. Bioinformatics. 2005;21: 4026-32.

Martin S, Zhang Z, Martino A, Faulon JL. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. Bioinformatics. 2007;23:866-74.

Mathelier A, Zhao X, Zhang A, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR. 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2013;42(Database issue):D142-7.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34: D108–10.

Menéndez P, Kourmpetis YA, ter Braak CJ, van Eeuwijk FA. Gene regulatory networks from multifactorial perturbations using Graphical Lasso: application to the DREAM4 challenge. PLoS One. 2010; 5(12):e14147.

Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol. 2007;79879.

Meyer P, Cokelaer T, Chandran D, Kim KH, Loh PR, Tucker G, et al. Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. BMC Syst Biol. 2014;8(1):13.

Mikheyev AS, Tin MM. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014;14:1097-102.

Miller MB, Bassler BL. Quorum sensing in bacteria. Annu Rev Microbiol. 2001;55:165-99.

Milnor J. On the concept of attractor. Commun Math Phys. 1985;99:177-95.

Mjolsness E, Mann T, Castaño R, Wold B. From co-expression to coregulation: An approach to inferring transcriptional regulation among gene classes from large-scale expression data. In: Solla SA, Leen TK, Müller KR (eds). Advances in neural information processing systems, Vol. 12 (pp 928-34). Cambridge, MA: MIT Press, 2000.

Montojo J, Zuberi K, Rodriguez H, Bader GD, Morris Q. GeneMANIA: Fast gene network construction and function prediction for Cytoscape. F1000Res. 2014 1;3:153.

Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques. 2008;45(1):81-94.

Mosher JJ, Bowman B, Bernberg EL, Shevchenko O, Kan J, Korlach J, et al. Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. J Microbiol Methods. 2014;104:59-60.

Mostafavi S, Debajyoti R, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology. 2008;9(Suppl 1):S4.

Mukherjee S, Sambarey A, Prashanthi K, Chandra N. Current trends in modeling host-pathogen interactions. WIREs Data Mining Knowl Discov. 2013;3: 109-28.

Murphy K, Mian S. Modelling gene expression data using dynamic Bayesian networks. Technical report. Berkeley, CA: Univ. of California at Berkeley, 1999.

Nakajima N, Akutsu T. Network completion for static gene expression data. Adv Bioinformatics. 2014; 2014:382452.

Nakajima N, Tamura T, Yamanishi Y, Horimoto K, Akutsu T. Network completion using dynamic programming and least-squares fitting. Sci World J. 2012;2012:957620.

Nelander S, Wang W, Nilsson B, She QB, Pratilas C, Rosen N, et al. Models from experiments: combinatorial drug perturbations of cancer cells. Mol Syst Biol. 2008;4:216.

Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio – the analysis and navigation of molecular networks. Bioinformatics. 2003;19:2155–215.

Olsen C, Bontempi G, Emmert-Streib F, Quackenbush J, Haibe-Kains B. Relevance of different prior knowledge sources for inferring gene interaction networks. Front Genet. 2014;5:177.

Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biology Direct. 2009;4(1):14-10.

Pallen MJ, Loman NJ, Penn CW. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr Opin Microbiol. 2010;13:625-31.

Perrin BE, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alché-Buc F. Gene networks inference using dynamic Bayesian networks. Bioinformatics. 2003;19 (Suppl 2):ii138-48.

Pinna A, Soranzo N, de la Fuente A. From knockouts to networks: establishing direct cause-effect relationships through graph analysis. PLoS One. 2010;5(10): e12912.

Pinna A, Heise S, Flassig RJ, de la Fuente A, Klamt S. Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: improved methods and their evaluation. BMC Syst Biol. 2013;7:73.

Plenge RM, Greenberg JD, Mangravite LM, Derry JM, Stahl EA, Coenen MJ, et al. Crowdsourcing genetic prediction of clinical utility in the Rheumatoid Arthritis Responder Challenge. Nat Genet. 2013;45: 468-9.

Prigogine I, Nicolis G. Biological order, structure and instabilities. Q Rev Biophys. 1971;4:107-48.

Priebe S, Menzel U, Zarse K, Groth M, Platzer M, Ristow M, et al. Extension of life span by impaired glucose metabolism in Caenorhabditis elegans is accompanied by structural rearrangements of the transcriptomic network. PLoS One. 2013;8(10):e77776.

Priebe S, Kreisel C, Horn F, Guthke R, Linde J. FungiFun2: a comprehensive online resource for systematic analysis of gene lists from fungal species. Bioinformatics. 2015;31:445-6.

Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, et al. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PLoS One. 2010;5(2):e9202.

Prill RJ, Saez-Rodriguez J, Alexopoulos LG, Sorger PK, Stolovitzky G. Crowdsourcing network inference: the DREAM predictive signaling network challenge. Sci Signal. 2011;4(189):mr7.

Rahmatallah Y, Emmert-Streib F, Glazko G. Comparative evaluation of gene set analysis approaches for RNA-Seq data. BMC Bioinformatics. 2014;15(1):397.

Rau A, Jaffrézic F, Foulley JL, Doerge RW. An empirical Bayesian method for estimating biological networks from temporal microarray data. Stat Appl Genet Mol Biol. 2010;9:article 9.

Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell. 2010;140:744-52. Erratum in: Cell. 2010;141:369. Kamburov A, Kaur M, MacPherson CR, Radovanovic A, Schwartz A [added].

Ruan J. A top-performing algorithm for the DREAM3 gene expression prediction challenge. PLoS One. 2010;5(2):e8944.

Ryll A, Bucher J, Bonin A, Bongard S, Gonçalves E, Saez-Rodriguez J, et al. A model integration approach linking signalling and gene-regulatory logic with kinetic metabolic models. Biosystems. 2014;124:26-38.

Samaga R, Klamt S. Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. Cell Commun Signal. 2013;11(1):43.

Sauer U, Heinemann M, Zamboni N. Genetics. Getting closer to the whole picture. Science. 2007; 316 (5824):550-1.

Schäfer J, Opgen-Rhein R, Strimmer K. Reverse engineering genetic networks using the GeneNet package. R News. 2006;5/6:50-3.

Schulze S, Henkel SG, Driesch D, Guthke R, Linde J. Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. Front Microbiol. 2015;6:65.

Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics. 2002;18:261-74.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genetics. 2014;15: 121–32.

Soneson D, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91.

Sorger PK. A reductionist's systems biology. Curr Opin Cell Biol. 2005;17:9–11.

Spieth C, Supper J, Streichert F, Speer N, Zell A. Jcell - a Java-based framework for inferring regulatory networks from time series data. Bioinformatics. 2006;22:2051-2.

Stolovitzky G, Monroe D, Califano A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. Ann N Y Acad Sci. 2007;1115:1-22.

Stolovitzky G, Prill RJ, Califano A. Lessons from the DREAM2 Challenges. Ann N Y Acad Sci. 2009; 1158:159-95.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):D447-52.

't Hoen PA, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013;31:1015-22.

Tekir SD, Cakır T, Ardıç E, Sayılırbaş AS, Konuk G, Konuk M, et al. PHISTO: pathogen-host interaction search tool. Bioinformatics. 2013;29:1357-8.

Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc Ser B. 1996;58:267-88.

Tierney L, Linde J, Müller S, Brunke S, Molina JC, Hube B, et al. An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. Front Microbiol. 2012;3:85.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expressionanalysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562-78.

Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13:146.

Ud-Dean SM, Gunawan R. Ensemble inference and inferability of gene regulatory networks. PLoS One. 2014;9(8):e103812.

van Someren EP, Wessels LF, Backer E, Reinders MJ. Genetic network modeling. Pharmacogenomics. 2002;3:507-25.

van Someren EP, Vaes BL, Steegenga WT, Sijbers AM, Dechering KJ, Reinders MJ. Least absolute regression network analysis of the murine osteoblast differentiation network. Bioinformatics. 2006;22:477-84.

Vignes M, Vandel J, Allouche D, Ramadan-Alban N, Cierco-Ayrolles C, Schiex T, et al. Gene regulatory network reconstruction using Bayesian networks, the Dantzig Selector, the Lasso and their meta-analysis. PLoS One. 2011;6(12):e29165.

Vlaic S, Schmidt-Heck W, Matz-Soja M, Marbach E, Linde J, Meyer-Baese A, et al. The extended TILAR approach: a novel tool for dynamic modeling of the transcription factor network regulating the adaption to in vitro cultivation of murine hepatocytes. BMC Syst Biol. 2012;6:147.

Wahde M, Hertz J. Coarse-grained reverse engineering of genetic regulatory networks. Biosystems. 2000;55(1-3):129-36.

Wang Y, Joshi T, Zhang XS, Xu D, Chen L. Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics. 2006;22:2413-20.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57-63.

Wang J, Zhang Y, Marian C, Ressom HW. Identification of aberrant pathways and network activities from high-throughput data. Brief Bioinform. 2012;13:406-19.

Weber M, Henkel SG, Vlaic S, Guthke R, van Zoelen EJ, Driesch D. Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0. BMC Syst Biol. 2013;7:1.

Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. Evaluation of methods for modelling transcription factor sequence specificity. Nat Biotechnol. 2013;31:126-34.

Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. Nat Rev Microbiol. 2012;10: 618-30.

Westra LR, Petreczky M, Peeters RLM. Identification of piecewise linear models of complex dynamic systems. ArXiv:1103.4756 [math.OC]. 2011.

Wu M, Chan C. Learning transcriptional regulation on a genome scale: a theoretical analysis based on gene expression data. Brief Bioinform. 2012;13:150-61.

Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011;39(Web Server issue):W316-22.

Ye X, Chu J, Zhuang Y, Zhang S. Multi-scale methodology: a key to deciphering systems biology. Front Biosci. 2005;10:961–5.

Yeung MK, Tegnér J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci U S A. 2002;99:6163-8.

Yeung KY, Dombek KM, Lo K, Mittler JE, Zhu J, Schadt EE, et al. Construction of regulatory networks using expression time-series data of a genotyped population. Proc Natl Acad Sci U S A. 2011;108:19436-41.

Yip KY, Alexander RP, Yan KK, Gerstein M. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. PLoS One. 2010;5(1):e8121.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.

Young WC, Raftery AE, Yeung KY. Fast Bayesian inference for gene regulatory networks using ScanBMA. BMC Syst Biol. 2014;8:47.

Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, et al. A comparative study of techniques for differential expression analysis on RNA-seq data. Plos One. 2014;9(8):e103207.

Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc Ser B. 2005; 67:301-20.

Zoppoli P, Morganella S, Ceccarelli M. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics. 2010;11:154.