frontiers in
GENETICS

REVIEW ARTICLE
published: 03 February 2012
doi: 10.3389/fgene.2012.00008

Statistical inference and reverse engineering of gene regulatory networks from observational expression data

Frank Emmert-Streib[1,2]*, Galina V. Glazko[3]*, Gökmen Altay[1,4,5] and Ricardo de Matos Simoes[1]

[1] *Computational Biology and Machine Learning Lab, School of Medicine, Dentistry and Biomedical Sciences, Center for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, UK*
[2] *Statistics and Computational Biology Laboratory, Li Ka Shing Centre, Cancer Research UK Cambridge Research Institute, University of Cambridge, Cambridge, UK*
[3] *Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA*
[4] *Department of Oncology, Cancer Research UK Cambridge Research Institute, University of Cambridge, Cambridge, UK*
[5] *Department of Electrical and Electronics Engineering, Bahcesehir University, Istanbul, Turkey*

Edited by:
Raya Khanin, Memorial Sloan-Kettering Cancer Center, USA

Reviewed by:
Raya Khanin, Memorial Sloan-Kettering Cancer Center, USA
Erik Larsson, Memorial Sloan-Kettering Cancer Center, USA

****Correspondence:***
Frank Emmert-Streib, Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK.
e-mail: v@bio-complexity.com;
Galina V. Glazko, Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA.
e-mail: gvglazko@uams.edu

In this paper, we present a systematic and conceptual overview of methods for inferring gene regulatory networks from observational gene expression data. Further, we discuss two classic approaches to infer causal structures and compare them with contemporary methods by providing a conceptual categorization thereof. We complement the above by surveying global and local evaluation measures for assessing the performance of inference algorithms.

Keywords: gene regulatory networks, statistical inference, reverse engineering, causal relations, directed acyclic graphs, Bayesian network, information-theory methods

1. INTRODUCTION

The purpose of this paper is to provide a systematic overview of methods used to estimate gene regulatory networks (GRN) from large-scale expression data. The inference of gene regulatory networks, which is sometimes also referred to as reverse engineering (Stolovitzky and Califano, 2007; Stolovitzky et al., 2009), is the process of estimating the direct physical (biochemical) interactions of a cellular system from data. That means one aims for identifying all molecular regulatory interactions among genes that are present in an organism to establish and maintain all required biological functions characterizing a certain physiological state of a cell. Depending on the data used for inferring the network, which, principally, may either come from DNA microarray, RNA-seq, proteomics or ChIP-chip experiments, or combinations thereof, the biological interpretation of an edge in these networks is dependent thereon. For expression data, inferred interactions may preferably indicate transcription regulation, but can also correspond to protein-protein interactions. Due to the causal character of these networks, which ensures a meaningful biological interpretation, the genome-wide inference of gene regulatory networks holds great promise in enhancing the understanding of normal cell physiology, and also complex pathological phenotypes (Barabási and Oltvai, 2004; Emmert-Streib, 2007; Schadt, 2009).

Due to the fact that this field is currently vastly expanding, this overview is inevitably incomplete. Instead of aiming to cover as many approaches as possible, we focus on conceptual clarity and methods for observational expression data. That means, we review statistical approaches from the literature we consider most important and show that they can be categorized nicely according to assumptions they make about the dynamic behavior of the data but also with respect to conceptual strategies they employ. In order to facilitate the understanding of the latter point we present also two seminal, and in the meanwhile classic, methods for the causal inference of networks and their theoretical foundations (Chow and Liu, 1968; Pearl, 1988; Spirtes et al., 1993). Model-based approaches based on, e.g., Boolean networks or differential equations (de Jong, 2002; Liu et al., 2008) are not covered in this paper. Also, supervised and semi-supervised network inference methods (Ernst et al., 2008; Mordelet and Vert, 2008; Cerulo et al., 2010) that require a training set will not be discussed. Further, we content ourselves with approaches for observational data (Rubin, 1974; Rosenbaum, 2002) neglecting methods that utilize interventional or perturbational data. In addition to the presentation of inference methods, we provide also an overview of global and local performance metrics frequently used to assess the inference abilities of such methods. Finally, we will emphasize and discuss the conceptual closeness of all current methods with respect to two classic methods (Chow and Liu, 1968; Pearl, 1988; Spirtes et al., 1993).

2. METHODS

The inference of gene networks from high-throughput data is a very complex and vastly expanding area triggered by the invention of measurement technologies. In order to provide a systematic discussion of the underlying principles we limit this review to observational, steady-state gene expression data, and consider correlation- and mutual information-based inference methods only, as visualized in **Figure 1**. These methods are representative of linear and non-linear methods. Principally, there are three fundamental levels of a molecular system as given by the central dogma of molecular biology (Crick, 1970), namely, the DNA, mRNA, and the protein level. Consequently, these levels imply sensible variables that can and should be measured to obtain information about the biological function of a cell. Specifically, one can distinguish between measurements that provide information about the protein-DNA binding (ChIP-chip, ChIP-seq), gene expression (DNA microarray, RNA-seq), and the protein-protein interaction (protein microarray) which can be used to infer various types of gene networks (Emmert-Streib and Glazko, 2011). We focus in this review on observational gene expression data because to date, this data type is predominating.

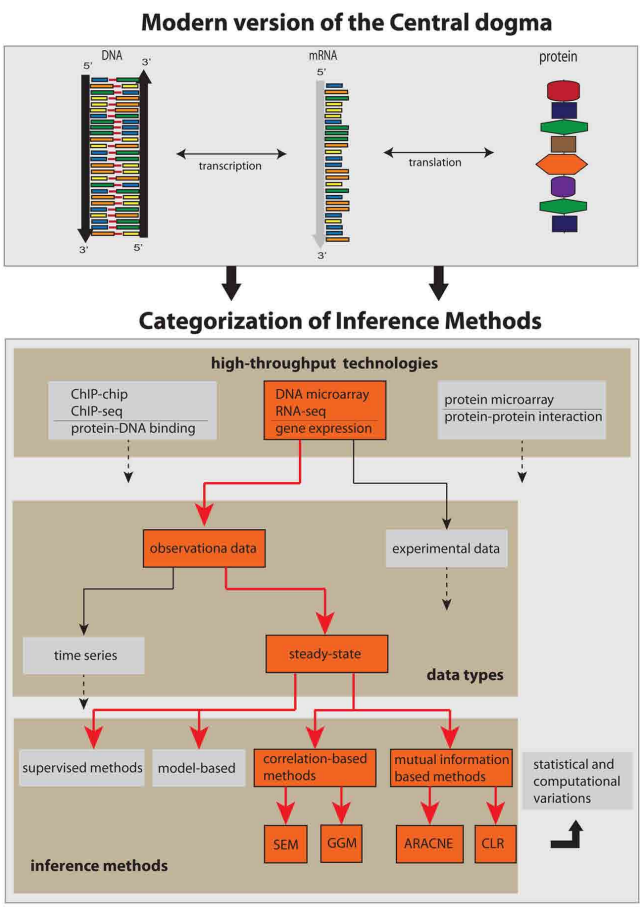

FIGURE 1 | Conceptual categorization of network inference methods and the biology, technology, and data types they depend on. The red arrows and boxes represent the focus of this review article.

The complexity of the network inference problem can be visualized with the help of **Figure 1**. There are two major factors that contribute to it. First, almost all components shown in **Figure 1** as represented by the boxes can be connected with each other. That means, they are not mutually exclusive but can be combined in a great variety. This concerns the integration of different high-throughput data, but also the combination of different data types or even methods. Second, any network inference method is subject to statistical and computational variations in form of technical modifications. This may relate to newly developed statistical estimators or optimization methods, or to the design of efficient algorithms or their parallelization on a computer cluster.

2.1. CLASSIC APPROACHES TO CAUSAL INFERENCE

We begin our presentation by some necessary preliminaries. Directed acyclic graphs (DAGs) are frequently employed to represent causal relations among variables (Wright, 1934; Verma and Pearl, 1990; Shipley, 2000). In such a graph, G, a directed edge from node X to Y means that X is the cause for Y. For this reason these networks are also called causal graphs. For the causal inference of network structures, the graph theoretical measure d-separation is key. It can be defined in the following way (Pearl, 1988; Verma and Pearl, 1990).

Definition 1. Two nodes X and Y are called d-separated by set S if and only if X, Y, and S are disjoint and every undirected path from X to Y is blocked by the nodes in set S.

If X is d-separated from Y by S we write $(X \perp\!\!\!\perp Y|S)$.

Definition 2. A path w is called d-separated or blocked by a set of nodes S if and only if either of the following two criteria hold:

1. on w is a node s which is no collider and $s \in S$
2. on w is a node s which is a collider and it holds $s \notin S$ and $de(s) \notin S$.

Here, $de(s)$ denotes the set of descendants of node s. A node Z is called a collider if the edge of the incoming as well as the edge of the outgoing link point toward Z. In **Figure 2** we clarify this notation. The top Figure shows a collider Z. For example, considering the path $X \rightarrow Z \leftarrow Y$, then from the definition of d-separation follows that X is d-separated from Y if conditioned on the empty set $S = \emptyset$, $(X \perp\!\!\!\perp Y|\emptyset)$, or simply $(X \perp\!\!\!\perp Y)$. However, if we condition on Z then $(X \not\perp\!\!\!\perp Y|Z)$, according to Definition 2. In this case Z unblocks the path between X and Y. In the middle and bottom panels in **Figure 2** two non-colliders (a chain and a fork) are shown. We want to notice that in general one node is sufficient to block a path.

A systematic connection between independence relations among variables V forming a DAG, and a process in form of a probability distribution P, can be given with the help of d-separation and Markov independencies. Formally, this connection is provided by an I-map (independency map). A DAG G is called an I-map of P if

$$(X \perp\!\!\!\perp Y|S)_G \Rightarrow (X \perp\!\!\!\perp Y|S)_P. \quad (1)$$

Here the subscript G refers to the structure represented by a DAG and the independence relations provided by d-separation,

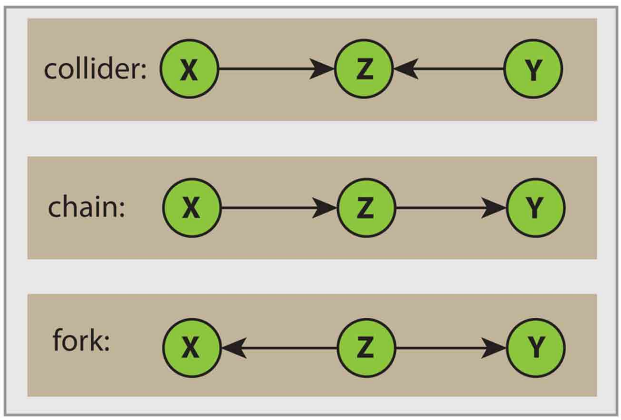

FIGURE 2 | Visualization of different path types. The orientation of edges with respect to the node Z allow the following distinction. Top: collider; middle: chain; bottom: fork.

and a "P" refers to the probability distribution whereas the independence relations corresponds to conditional Markov independencies. From equation (1) follows the definition of a Bayesian network (Pearl, 1988).

Definition 3. *Bayesian network* Given a probability distribution P on a set of random variables V. A DAG, $G = (V, E)$, is called a Bayesian network of P if and only if G is a minimal I-map of P.

Here minimal I-map means that if any edge in the DAG is deleted, G is no longer an I-map of P. There are alternative definitions of a Bayesian network (Pearl, 2000), however, the definition given above emphasizes its connection to d-separation best. This is important to emphasize because Bayesian networks are first of all about independence relations, and not about specific probability distributions.

In our opinion, there are basically two principle approaches in the literature that are relevant for our contextual problem, which *proof* mathematically that, under certain conditions, they are capable of systematically inferring causal relations. The first principle approach, which is based on *d-separation*, has been independently developed by two group; Verma and Pearl (1991) and Spirtes et al. (1993). The second principle approach is from Chow and Liu (1968). Certainly, there are various variations of these two approaches, however, in the following, we focus only on the essential principles of both methods.

The first algorithm we discuss proven to reconstruct a causal structure is the *inductive causation* (IC) algorithm. For simplicity, we assume *causal sufficiency*, which means that latent variables are absent. The IC algorithm starts with the construction of an undirected dependency graph (UDG) by testing exhaustively for the independence of variable X from Y given a set S_{xy}. Exhaustively means, for any set S_{xy} that can be formed among the available variables. We denote this by $(X \perp\!\!\!\perp Y \mid S_{xy})_P$ emphasizing explicitly that this independence is with respect to an underlying distribution P.

The basic principle on which the IC algorithm is based on, is a connection between the d-separation relations of a DAG G and independence relations of a probability distribution P consistent with the structure of the DAG G. The crucial point is that the distribution P and its independence relations are not given (known) but they need to be estimated from data, generated from the distribution P. Schematically, this is outlined in **Figure 3**. Here, the red line corresponds to the experimentally generation of the data, e.g., by means of DNA microarrays, and the blue line corresponds to the theoretical, however, unobservable (that's why this line is dashed), independence relations present in the DAG G that can be expressed by the concept of d-separation. Based on the data, an inference method aims for estimating these independence relations statistically. The gray box marks the range of the inference method.

The second algorithm proven to reconstruct a causal structure is from Chow and Liu (1968). It is based on the MWST (maximum weight spanning tree) algorithm, which connects node pairs in a way that the resulting network structure forms a tree with a minimal sum of edge weights. It is of interest to note that the proof of the MWST algorithm (see for example Pearl, 1988) is based on the *data processing inequality* (DPI; Cover and Thomas, 1991), which will be discussed later in the paper for ARACNE.

2.2. CATEGORIZATION OF METHODS

In this section we present an overview and a categorization of methods that have been *specifically* introduced to infer gene regulatory networks from observational gene expression data. Previously, there have been other reviews that attempted to do this (Werhli et al., 2006; Bansal et al., 2007; Margolin and Califano, 2007; Markowetz and Spang, 2007; Hache et al., 2009; Olsen et al., 2009; de Smet and Marchal, 2010; Penfold and Wild, 2011), however, their emphasis and conceptualization deviates from ours.

2.2.1. Correlation-based estimation methods

2.2.1.1. Co-expression networks. There are two principally different ways to construct co-expression networks from microarray data. One approach follows a hard- and the other a soft-thresholding of correlation coefficients (Zhou et al., 2002; Horvath and Dong, 2008). In the following $r_{ij} = |cor(i,j)|$ is the absolute value of the Pearson correlation coefficient. In the case of hard-thresholding, the correlation coefficient r_{ij} between gene i and j is assessed with respect to a threshold τ and both genes are connected by an edge, $A_{ij} = A_{ji} = 1$, if $r_{ij} \geq \tau$, otherwise the genes remain unconnected, $A_{ij} = A_{ji} = 0$. The threshold or cut-off parameter τ can be obtained from randomization of the data allowing the assessment of statistical significance (Carter et al., 2004). This results in an undirected, unweighted network.

For the soft-thresholding two types of adjacency functions are frequently used (Zhang and Horvath, 2005). The sigmoid function,

$$A_{ij} = \frac{1}{1 + exp(-\alpha(r_{ij} - \tau_0)}, \tag{2}$$

and the power adjacency function,

$$A_{ij} = |r_{ij}|^{\beta}. \tag{3}$$

Both types of adjacency functions lead to undirected but weighted networks. In order to choose the above parameters appropriately

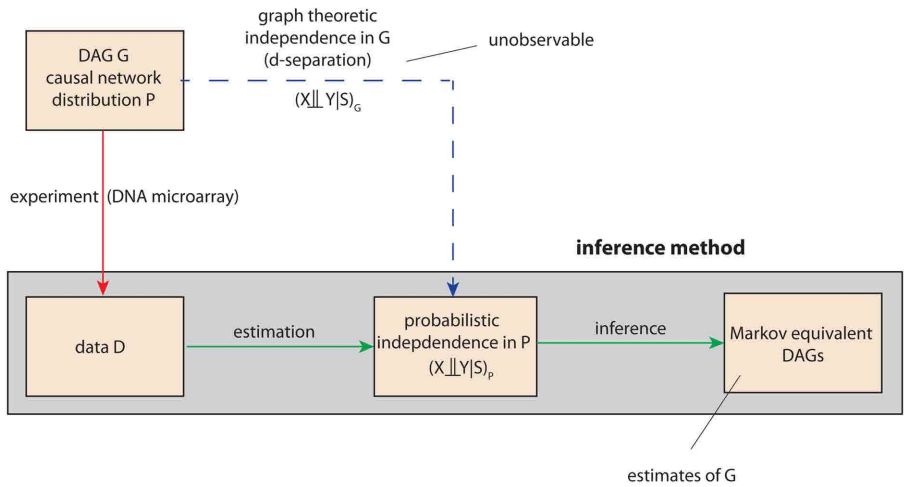

FIGURE 3 | Schematic visualization of the importance of the d-separation concept in the inferential task to recover the structure of a DAG *G*. The red, green, and blue lines refer to experimental, inferential, or theoretical (unobservable) steps. For notational simplicity, we use also in the inferential step the symbol P instead of $\hat{P}$ to indicate an estimate of P.

Zhang and Horvath (2005) suggested to adjust them in a way that the resulting network has approximately a scale-free degree distribution.

We would like to remark that the purpose for the construction of co-expression networks is different to all other methods discussed in this paper. Co-expression networks serve as means to explore the functionality of genes on a systems level (Zhang and Horvath, 2005) and do not aim to be causal representations of regulatory networks. Nevertheless, we included them in this review to point out that also networks that are not causal can be very useful in gaining biological understanding and insights.

2.2.1.2. Asymmetric-N. Asymmetric-N is an algorithm that takes the fact into account that biological networks contain hubs. It is a modified version of Symmetric-N (Agrawal, 2002) which was designed for the construction of co-expression networks (Chen et al., 2008a). The two-step algorithm of Agrawal (2002), Symmetric-N, utilizes the N-nearest-neighbor concept. In a first step, for each node in the network its correlation to other nodes is calculated and sorted in descending order. In a second step, each pair of nodes is evaluated one by one and if for a pair of nodes both are in the corresponding N-nearest-neighbors list, a connection between them is included. Otherwise, they are not connected. Here N is a pre-defined cut-off value for the number of neighbor nodes to be considered (Agrawal, 2002). It is demonstrated in Agrawal (2002) that the inferred network has a scale-free topology.

Instead of using N neighbors for all nodes as potential connections, a higher number N_C is assigned to some core nodes (e.g., TFs) and a smaller number N_P is assigned to periphery nodes (e.g., non-TFs). Since the numbers of potential neighbors are different for core and peripheral nodes, the algorithm was called Asymmetric-N (Chen et al., 2008a). Asymmetric-N employs the intuition that transcription factors (TFs) are more frequently connected to other genes than regulated genes. An interesting result found in Chen et al. (2008a) is that Asymmetric-N performed better than Bayesian networks for small sample sizes.

2.2.1.3. SEM (structural equation model). Xiong et al. (2004) use a structural equation model (SEM; Jöreskog, 1973; Bollen, 1989) to reconstruct gene networks. Structural equation models combine path analysis (Wright, 1921, 1934) and confirmatory factor analysis (Jöreskog, 1969; Brown, 2006).

Xiong et al.'s (2004) approach starts by assuming that the expression level of genes, X, can be described linearly by

$$X = BX + \Gamma\xi + \varepsilon. \tag{4}$$

Here X is a vector of p endogenous and ξ a vector of q exogenous variables, and B and Γ are $p \times p$ respectively $p \times q$ coupling matrices. The last term in equation (4), ε, represents noise. The structure of the regulatory gene network is coded by the entries of the coupling matrices B and Γ. The structural equation in equation (4) is learned in two steps. First, assuming the structure of the network is known the coupling parameters are estimated via maximum likelihood optimizing the distance between the $(p+q) \times (p+q)$ covariance matrix $\Sigma(\Theta)$ and the sample covariance matrix S estimated from the data. Here, $\Theta = f(B, \Gamma, \Phi, \Psi)$, is a function of the parameters of the SEM and the covariances of ξ (Φ) and ε (Ψ) that can be expressed analytically. The second step consists in finding the network structure (of B and Γ) by an optimization method whereas the different models are assessed by Akaike's information criterion (AIC).

2.2.1.4. Low-order partial correlation. Partial correlations of low-order have been employed in de la Fuente et al. (2004), Magwene and Kim (2004), Wille and Bühlmann (2006). This approach is capable to infer undirected networks only, because a correlation, of any order, is symmetric in its arguments. The principle working mechanism of this approach is as follows.

Starting with a fully connected network, interactions between genes are iteratively excluded by moving toward an increasing order of the correlation coefficient. More specifically, starting from a fully connected adjacency matrix A, one calculates in the first step all pair-wise correlation coefficients, $\rho_{ij} = \rho(X_i, X_j)$. If $\rho_{ij} = 0$ the connection between gene X_i and X_j is deleted, $A_{ij} = A_{ji} = 0$. Inthe second step all partial correlation coefficients of first-order, $\rho_{ij|k} = \rho_{X_i,X_j|X_k}$, are calculated for all triplets of genes with $\rho_{ij} \neq 0$. Now, if there is at least one gene X_k for which $\rho_{ij}|k = 0$ holds the connection between gene X_i and X_j is deleted, $A_{ij} = A_{ji} = 0$. Continuation to higher orders is analogously. The exclusion of interactions is based on testing the null hypothesis $\rho_{ij} = 0$ against the alternative hypothesis $\rho_{ij} \neq 0$ for all orders of the (partial) correlation coefficients. In (de la Fuente et al., 2004) this approach was applied up to order three. The resulting network from this procedure is frequently called an *undirected dependency graph* (UDG; Shipley, 2000). An implementation of this approach is called ParCorA (de la Fuente et al., 2004).

2.2.1.5. GGM (graphical Gaussian model). GGM, also known as covariance selection model, concentration graph, or Markov random field (Dempster, 1972; Whittaker, 1990; Koller and Friedman, 2009), is a graphical model which assumes that all variables are distributed according to a multivariate normal distribution with a specific structure of the inverse of the covariance matrix, $\Omega = \Sigma^{-1}$, also called precision or concentration matrix. Network inference methods based on GGM make use of the relation,

$$\rho_{ij|V\setminus\{ij\}} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}, \tag{5}$$

connecting the partial correlation of full-order with the elements of Ω, $w_{ij} \in \Omega$. The partial correlation is of full-order (with respect to the number of genes) because $V\setminus\{ij\}$ is the set of all genes excluding i and j, i.e., the largest possible set of genes not considering i and j. In case of a vanishing partial correlation in equation (5) one can write

$$\left(X_i \perp\!\!\!\perp X_j | V\setminus\{ij\}\right) = 0. \tag{6}$$

From this conditional independence relation, the principle way to infer a network structure from GGM becomes apparent estimating a network in the following way. If $\rho_{ij}|V\setminus\{ij\} \neq 0$, according to a hypothesis test, we include an edge, $A_{ij} = A_{ji} = 1$, otherwise there is no edge between i and j.

Several approaches have been made to infer gene regulatory networks based on GGM (Wille et al., 2004; Schäfer and Strimmer, 2005; Li and Gui, 2006). These methods differ in the way the inverse of the covariance matrix, Σ^{-1}, is estimated and in the statistical tests employed to define significance. The reason for these technical variants comes from a variety of problems. First, if the number of samples is smaller than the number of genes, which is typically the case for genomics data, the sample covariance matrix is not positive definite and, hence, not invertible. Another problem is caused by the small samples size problem (Schäfer and Strimmer, 2005).

2.2.1.6. BN (Bayesian networks). Bayesian networks allow identifying a DAG structure of a network. BN were among the first methods that have been applied to expression data to infer GRN (Friedman et al., 2000; Hartemink et al., 2001). In order to overcome the problem that only acyclic networks can be inferred by BN, which would be a severe limitation considering the fact that real biological networks contain many feedback loops and are, hence, cyclic, a modified method called *dynamic Bayesian network* (DBN) is used instead. Briefly, DBN unfold cyclic processes by mapping them onto a sequence of acyclic events, which can then be analyzed with a BN (Dean and Kanazawa, 1990; Koller and Friedman, 2009). For the inference of GRN, this approach has been applied in Husmeier (2003), Perrin et al. (2003), Zou and Conzen (2005). The variations in these approaches provide different methods for the *structure learning* problem, which is the integral part of a BN learned from data. Principally, a major practical problem of a BN is that the computational complexity of algorithms to learn their structure has been shown to be NP-hard for score-based approaches (Chickering et al., 1995; Chickering, 1996). Hence, BN or DBN can be either only applied to quite small networks or heuristic approximation methods have to be used that separate the score into tractable units that can be optimized by locally constraint search techniques making the computational complexity manageable (Friedman et al., 1999, 2000).

In **Figure 6**, we categorize BN as correlation-based methods because the simplest statistical realization of a BN is obtained by estimating $(X \perp\!\!\!\perp Y|S)_P$, see section 1, by means of partial correlation coefficients (Spirtes et al., 1993; Pearl, 2000). However, also non-linear approaches are possible.

2.2.2. Information-theory based estimation methods

In the following, we restrict our discussion to the working mechanisms of the discussed procedures. However, we want to emphasize that also for these methods the employed statistical estimators are of importance (Beirlant et al., 1997; Steuer et al., 2002; Daub et al., 2004; Kraskov et al., 2004; Khan et al., 2007).

2.2.2.1. A. Mutual information-based. *RN (relevance networks).* The principle idea of RN (Butte and Kohane, 2000) is to compute all mutual information (MI) values for all pairs of genes, for a given data set, and declare mutual information values as relevant if their corresponding value is larger than a given threshold I_0. The resulting network is constructed based on this threshold by including an edge between two genes in the respective adjacency matrix of the network, $A_{ij} = A_{ji} = 1$, if $I_{ij} > I_0$, otherwise no edge is included between i and j. In Butte and Kohane (2000) the threshold I_0 was found by randomization of the expression data set. From this randomization, mutual information values were re-calculated from which a reference distribution of mutual information values, resembling a null-distribution, was obtained. Based on this reference distribution the threshold I_0 was obtained by heuristic arguments. For this reason, mutual information values that are larger than I_0 are called relevant but cannot necessarily be called statistically significant.

ARACNE (algorithm for the reconstruction of accurate cellular networks). ARACNE (Basso et al., 2005; Margolin et al., 2006) is

similar to RN because it also uses mutual information and a randomization of the data to obtain a threshold I_0 allowing to declare mutual information values significant if $I_{ij} > I_0$[1]. If I_{ij} is found to be significant, than an edge is included in the corresponding adjacency matrix between gene i and j, $A_{ij} = A_{ji} = 1$, otherwise no edge is included. However, in contrast to RN, ARACNE performs a second step testing all gene-triplets (three genes with mutual information values larger than I_0) such that, for each triplet (ijk), the edge corresponding to the lowest mutual information value $I_1 = I_{i'j'}$, with $(i'j') = \text{argmin}\{I_{ij}, I_{jk}, I_{ik}\}$, is eliminated from the adjacency matrix, if it is smaller than the second smallest MI value I_2 multiplied by a factor, i.e.,

$$A_{i'j'} = A_{j'i'} = \begin{cases} 0 & I_{i'j'} \leq I_2(1-\epsilon) \\ 1 & \text{otherwise.} \end{cases} \tag{7}$$

Here $0 \leq \epsilon \leq 1$. The introduction of this step has been motivated by the so called *data processing inequality* (DPI; Cover and Thomas, 1991). The DPI is a relation between mutual information values which means loosely that a post-processing of data cannot increase its information content. Specifically, one can show (Cover and Thomas, 1991) that the DPI for the following relation between the three random variables,

$$X \rightarrow Y \rightarrow Z, \tag{8}$$

implies that $I(X,Z) \leq I(X,Y)$. Due to the fact that the criteria in equation (8) is for $\epsilon > 0$ less stringent than the DPI [equation (9)], ϵ is called *tolerance parameter*.

To ensure that the application of equation (8) results in an unique solution, independent of the order the triples have been selected, the procedure starts by listing all possible triplets in the network G. Then all of these triplets are tested. Hence, the results of these tests have no influence on subsequent tests of triplets. At the end of the second step, the resulting network represents the final result.

ARACNE employs two parameters, I_0 and ϵ. The cut-off parameter I_0 is determined by a resampling method estimating the distribution of the null hypothesis corresponding to a vanishing mutual information. This allows to assign p-values to mutual information values. In contrast, the DPI is not directly connected to statistical inference, but serves as a filtering step. Optimal values of ϵ are found from simulation studies that allow a comparison with the underlying true network. Hence, I_0 is found in an unsupervised and ϵ in a supervised way of learning.

CLR (context likelihood of relatedness). CLR (Faith et al., 2007) is also an extension of RN. It starts by estimating the pair-wise mutual information values for all genes. Then, CLR estimates the statistical likelihood of each MI value I_{ij}, for a particular pair of genes (ij), by comparing this MI value to a background. Specifically, for each gene pair (ij) two z-scores are obtained, one for gene i and one for gene j, by comparing the mutual information value I_{ij} with gene specific distributions, p_i and p_j. Here the two distributions p_i and p_j correspond to the distributions of mutual information values related to gene i ($\{I_{ik} | k \in V\}$) and gene j ($\{I_{jk} | k \in V\}$). By making a normality assumption about these distributions, corresponding z-scores, z_i and z_j, can be obtained from which the joint likelihood measure $\overline{z_{ij}} = \sqrt{z_i^2 + z_j^2}$ is calculated. In contrast to RN and ARACNE, which employ a global threshold I_0 for each mutual information value correspondingly pair of genes, CLR estimates *individual* thresholds by considering an *individual* background for each pair of genes. This procedure can be seen as a first attempt to take the network context of gene pairs into account.

C3NET (conservative causal core). C3NET consists of two main steps (Altay and Emmert-Streib, 2010a). The first step is for the elimination of non-significant edges, whereas the second step selects for each gene the edge among the remaining ones with maximum mutual in formation value. The first step is similar to RN (Butte and Kohane, 2000), ARACNE (Margolin et al., 2006), or CLR (Faith et al., 2007) essential for eliminating non-significant links, according to a chosen significance level α, between gene pairs. In the second step, the most significant link for each gene is selected. This link corresponds also to the highest MI value among the neighbor edges for each gene. This implies that the highest possible number of edges that can be inferred by C3NET is equal to the number of genes under consideration. This number can decrease for several reasons. For example, when two genes have the same edge with maximum MI value. In this case, the same edge would be chosen by both genes to be included in the network. However, if an edge is already present another inclusion does not lead to an additional edge. Another case corresponds to the situation when a gene does not have significant edges at all. In this case, apparently, no edge can be included in the network. Since C3NET employs MI values as test statistics among genes, there is no directional information that can be inferred thereof. Hence, the resulting network is undirected and unweighted. For a detailed explanation of C3NET and technical details, the reader is referred to Altay and Emmert-Streib (2010a, 2011), de Matos Simoes and Emmert-Streib (2011).

2.2.2.2. B. Extensions of mutual information and conditional mutual information.

SA-CLR (synergy augmented CLR). As indicated by its name, SA-CLR is based on CLR but includes synergistic effects between genes (Anastassiou, 2007; Watkinson et al., 2009). Synergy,

$$\text{Syn}\,(X, Y; Z) = I\,(X, Y; Z) - (I(X; Z) + I(Y; Z))\,, \tag{9}$$

is defined as the difference between the generalized two-way mutual information,

$$I(X, Y; Z) = \mathbb{E}_{p_{(x,y,z)}} log \frac{p(x, y, z)}{p(x, y)p(z)}, \tag{10}$$

and the mutual information values for two random variables. A visualization of synergy is shown in **Figure 4**. In this figure, the different intersections of the three variables X, Y, and Z are labeled by lower case letters, whereas "e" corresponds to the synergy,

[1]The details of the randomization are different to that used by RN allowing now to make statistical statements.

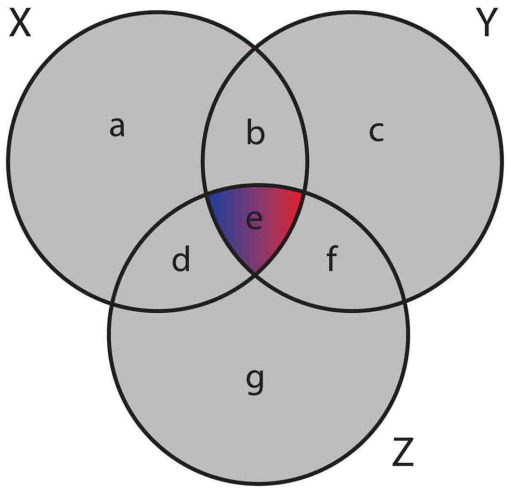

FIGURE 4 | Generalized Venn diagram showing the synergy Syn(*X,Y;Z*) = − *I*(*X;Y;Z*). The synergy is shown in red and blue because it can assume positive as well as negative values.

Syn(X,Y;Z). Using the definitions of the occurring mutual information values one obtains a simplified form of the synergy, given by

$$\begin{aligned}
\text{Syn}(X, Y; Z) &= \mathbb{E}_{p(x,y,z)} log \frac{p(x,y,z)}{p(x,y)p(z)} - && (11)\\
&\quad \mathbb{E}_{p(x,z)} log \frac{p(x,z)}{p(x)p(z)} - \mathbb{E}_{p(y,z)} log \frac{p(y,z)}{p(y)p(z)} && (12)\\
&= \mathbb{E}_{p(x,y,z)} log \frac{p(x,y,z)}{p(x,y)p(z)} \frac{p(x)p(z)}{p(x,z)} \frac{p(y)p(z)}{p(y,z)} && (13)\\
&= \mathbb{E}_{p(x,y,z)} log \frac{p(x)p(y)p(z)p(x,y,z)}{p(x,y)p(y,z)p(y,z)}\\
&= -I(X;Y;Z) && (14)
\end{aligned}$$

which is the negative of the three-dimensional interaction information (Watkinson et al., 2009). We want to remark that $I(X; Y; Z)$ can assume positive as well as negative values (McGill, 1954). For this reason the intersection "e" in **Figure 4** is shown in blue and red to emphasize this. Due to the fact that a Venn diagram visualizes logical relations between variables **Figure 4** is no Venn diagram in the strict sense but a generalization thereof, allowing also to express negative values by including colors in the representation. The idea of SA-CLR consists in utilizing the synergy given in equation (10) in the following way,

$$S(X, Y) = \max_{\substack{Z \\ \text{for } X \neq Z, Y \neq Z \\ I(X;Z) < I(X;Y) \\ I(X;Z) < I(Y;Z)}} \left(-I(X;Y;Z)\right). \quad (15)$$

This term is called *synergistic regulation index* (SRI; Watkinson et al., 2009). The interpretation of SRI is that for two given genes, X and Y, we are searching a third one, Z, that maximizes the synergy Syn(X, Y; Z) obeying the above constraints. This gene, Z, can be be arbitrarily chosen among all available genes, besides X and Y. The two constraints that involve mutual information, are chosen assuming that gene X and Z regulate Y. In this case, the mutual information between the two regulators, X and Z, should be smallest, motivating the constraints. Interestingly, these constraints introduce an asymmetry in X and Y[2] making it possible to speak about regulator and modulator genes. In the case of a positive synergy, a significant high value of SRI may indicate that gene X regulates genes Y (Watkinson et al., 2009).

In order to detect also non-cooperative effects, the measure actually used by SA-CLR is $I(X, Y) + S(X, Y)$, the mutual information between two genes plus their *synergistic regulation index*. The significance of $I(X, Y) + S(X, Y)$ is detected in a similar way as described above for CLR.

It is interesting to note that

$$I(X; Y) + S(X, Y) = I(X; Y|Z)|_{Syn(X,Y;Z) > 0} \quad (16)$$

holds formally but is not identical to the (general) conditional mutual information $I(X; Y|Z)$ because the synergy for X, Y, and Z has to be positive. To make this clear, we included this constraint visibly in equation (20). This allows SA-CLR, in contrast to CLR, to obtain a directed network by assigning an edge from gene X to Y if $I(X; Y) + S(X, Y)$ is significant.

MRNET (maximum relevance, minimum redundance). MRNET (Meyer et al., 2007) is an iterative algorithm that identifies potential interaction partners of a target gene Y that maximize a scoring function. Its working mechanism is as follows,

$$X_j^s = \underset{X_j \in V \setminus S}{\text{argmax}} (s_j) \quad (17)$$

$$s_j = I(X_j; Y) - \frac{1}{|S|} \sum_{X_k \in S} I(X_j; X_k). \quad (18)$$

Whenever a gene, X_j, is found with a score that maximizes equation (21) and s_j is above a threshold, s_0, then this gene is added to the set S. The algorithm is iterated until no further gene can be found that would pass the threshold test. The basic idea of MRNET is to find interaction partners for Y that are of maximal relevance [first term in equation (22)] for Y, but introduce a minimum redundancy [second term in equation (22)] with respect to the already found interaction partners in the set S. Starting with a fully connected, undirected network among all genes, MRNET reduces successively edges between Y and VS, which have not maximized the score in equation 22.

We want to remark that the score s_j can be approximated by the difference between two mutual information values,

$$s_j \approx I(X_j; Y) - I(X_j; A), \quad (19)$$

where the (auxiliary) random variable, A, represents the influence of the variables in the set S. Equation (23) is visualized in the left

[2]We want to remark that $I(X; Y; Z)$ is symmetric in the three variables.

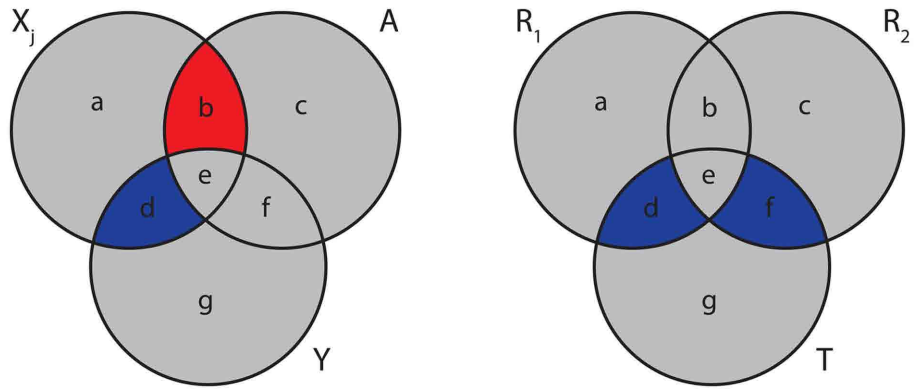

FIGURE 5 | Generalized Venn diagrams of the information captured by different measures. Left: score s_j [equation (23)] for MRNET. Right: the MI_3 measure of the MI3 algorithm. Blue corresponds to positive and red to a negative values.

Figure 5. Here, $I(X_j; Y)$ corresponds to "d" and $I(X_j; A)$ to "b." Due to the fact that the mutual information is always positive, $I(X_j; Y)$ adds a positive and $I(X_j; A)$ a negative value to s_j, indicated by blue and red colors. It has been argued in (Meyer et al., 2007) that due to the maximization of equation (22), MRNET approximates the conditional mutual information $I(X_j;Y|S)$. This is also motivated by **Figure 5**.

MI3 (mutual information 3). The MI3 algorithm (Luo et al., 2008) uses three-way mutual information for the inference, hypothesizing that gene regulation commonly involves more than one regulator genes. The value of the measure MI_3 is defined as,

$$MI_3\,(T, R_1, R_2) = 2I\,(T; R_1, R_2) - (I(T; R_1) + I(T; R_2))\,, \quad (20)$$

which equals the difference of mutual informations between the target gene and the two regulators and the target gene with one of the regulators. Alternatively, MI_3 can also be written as

$$\begin{aligned} MI_3(T, R_1, R_2) &= I(T; R_1|R_2) + I(T; R_2|R_1) && (21)\\ &= I(T; R_1, R_2) + Syn(T, R_1, R_2) && (22)\\ &= I(T; R_1, R_2) - I(T; R_1; R_2). && (23) \end{aligned}$$

Equation (25) is the sum of two conditional mutual informations between the target gene and a regulator given the other regulator, which emphasizes the three-way nature of this measure. Interestingly, MI_3 has also a simple relation to synergy, as shown in equation (26). This follows directly from equations (10 and 24).

In the right **Figure 5**, we show a visualization of MI_3. Here, the shown intersections result from the correspondence of $I(T; R_1, R_2)$ to "d + e + f," $I(T; R_1)$ to "d + e," and $I(T; R_2)$ to "e + f." Subtraction of these terms according to equation (24) results in "d + f." Note that $I(T; R_1, R_2)$ is counted twice.

The MI3 algorithm learns gene regulatory networks in two steps: first, local regulatory networks consisting of only three genes (T, R_1, and R_2) are learned. Starting from a given gene, T, two regulator genes are searched which maximize MI3 (T, R_1, R_2). As a result, directed edges between R_1 and T and R_2 and T are added constituting a small regulatory network. Second, the local regulatory networks learned in step one are assembled. To ensure that the resulting network is acyclic, there may be a need to remove edges forming cycles in the assembled network. This is solved heuristically, by identifying all local three-gene networks that contribute to a cycle and elimination of all edges of them from the overall network, except from the one with the highest MI_3 value. Overall, the MI3 algorithm aims to learn the optimal two-parent causal model for each target variable in the form R1 → T and R2 → T.

CMI (conditional mutual information). Soranzo et al. (2007) use only the conditional mutual information (CMI) to infer regulatory networks. They estimate all conditional mutual information values $I(X_i; X_j | X_k)$ between triplets of genes. Starting from a fully connected network, they successively remove edges $A_{ij} = A_{ji} = 0$ if $I(X_i; X_j | X_k) = 0$ for at least one gene X_k, according to a threshold condition.

MI-CMI. The MI-CMI algorithm (Liang and Wang, 2008) uses both mutual information and conditional mutual information to estimate the network. Starting from a fully unconnected network G, the algorithm consists of three steps. First, all pairs of mutual information values $I(X_i; X_j)$ are estimated. If $I(X_i; X_j) \geq I_0$, then an edge is included, $A_{ij} = A_{ji} = 1$. Second, for all triplets (X_i, X_j, X_k) with $I(X_i; X_j) \geq I_0$, $I(X_i; X_k) \geq I_0$, and $I(X_j; X_k) \geq I_0$ the conditional mutual information for all combinations of $I(X_i, X_j, X_k)$ is estimated. Then based on heuristic rules, is it decided which edges in G are likely to result from indirect regulations (interactions). These edges will be removed from G. Third, for all triplets (X_i, X_j, X_k) with $I(X_i; X_j) < I_0$, $I(X_i; X_k) < I_0$, and $I(X_j; X_k) < I_0$ the conditional mutual information for all combinations of (X_i; X_j; X_k) is estimated. Again, following some heuristic rules, edges that are likely the result from interactive regulations, which could not be detected by mutual information, are included in G. Special emphasize was given to the used statistical estimators, for the mutual information and conditional mutual information values.

Mutual information in combination with conditional mutual information was also used in Zhao et al. (2008), Zhang et al. (2011).

However, in contrast to (Liang and Wang, 2008), these approaches are conceptually closer to the PC algorithm (Spirtes et al., 1993; Shipley, 2000).

2.3. CATEGORIZATION FROM A DYNAMICAL PERSPECTIVE

In **Figure 6** we present an overview of inference methods to infer regulatory networks, as discussed in the previous section. We separated these methods in two main groups, depending on if the method is correlation- (pink) or information-based (orange). Within these two groups, there are various ways for further subdivision, however, we just distinguish between basic forms and higher order extensions – as indicated by darker colors (see **Figure 6**).

2.4. MULTIPLE TESTING CORRECTIONS

Due to the fact that all methods presented above involve many hypotheses that are tested, one needs to apply a multiple hypothesis correction method (Lehmann and Romano, 2005; Dudoit and van der Laan, 2007). From the investigation of multiple testing corrections it is known that the presence of a correlation structure in the data can lead to severe problems that counteracts an efficient control of an error measure, e.g., of the false discovery rate (FDR) or the family-wise error (Benjamini and Yekutieli, 2001; Storey and Tibshirani, 2003; Dudoit et al., 2008). Unfortunately, in the specific context of the inference of regulatory networks, multiple testing corrections have not been studied well. For this reason, without conducting in depth investigations of this problem one should apply a conservative rather than a more liberal correction method. Hence, one can apply a Bonferroni correction (Bonferroni, 1936; Dudoit and van der Laan, 2007).

3. EVALUATION MEASURES

In order to assess the performance of inference methods several measures have been suggested. In the following, we present three different types of such measures: (1) General statistical measures, (2) Ontology-based measures, (3) Network-based measures. On overview of these different measures is given in **Figure 7**.

3.1. GENERAL STATISTICAL MEASURES

The most widely used statistical measures are based on,

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{24}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{25}$$

$$\text{complementary sensitivity} = 1 - \text{sensitivity} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{26}$$

$$\text{precision} = \text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{27}$$

$$\text{recall} = \text{R} = \text{sensitivity} \tag{28}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{29}$$

obtained by comparison of a inferred (predicted) network with the true network underlying the data. From the measures listed above, three pair-wise combinations thereof are frequently used to assess the performance of an algorithm. The first measure is the area under the curve for the *receiver operator characteristics* (AUC-ROC; Fawcett, 2006). The ROC curve represents the sensitivity as function of the complementary specificity obtained by using various threshold values $\theta \in \Theta$, instead of one specific, the algorithm depends on[3]. This leads to a θ-dependence of all quantities listed above and, hence, allows to obtain a functional behavior among these measures. The second measure is the area under the precision-recall curve (AUC-PR), obtained similarly as AUC-ROC, and the third is the F-score,

$$\text{F} = 2\frac{\text{PR}}{\text{P} + \text{R}}, \tag{30}$$

also called F_1 because it is a special form of,

$$\text{F}_\beta = (1 - \beta^2)\frac{\text{PR}}{\beta^2(\text{P} + \text{R})}. \tag{31}$$

We want to emphasize that all measures presented above are general statistical measures used in statistics and data analysis. None of them is specific to our problem under consideration, namely, the inference of regulatory networks from expression data. In other words, none of these measures utilizes either biological or network specific information in any form. Further, each of these general statistical measures are global error measures because they evaluate the network inference performance as a whole, represented by a scalar value. As a consequence thereof, it is implicitly assumed that the inference process is homogeneous, i.e., each interaction should have about the same true positive rate, because otherwise it would be implausible to summarize the inference performance by just one value, e.g., a F-score. However, this is not justified, as we will discuss below.

In the following, we present extensions of general statistical measures for both types of information.

3.2. ONTOLOGY-BASED MEASURES

An evaluation strategy utilizing biological information to assess the performance of an inference method tries to quantify the *biological relevance* of the inferred network. In general, it is assumed that genes in a gene regulatory network are preferentially linked to genes involved in similar biological processes (Wolfe et al., 2005). There are several publicly available resources of biological knowledge which can be used to test whether this holds true, for example, the Gene Ontology database (GO; Ashburner et al., 2000), or KEGG (Kanehisa and Goto, 2000). There are also several curated organism-specific knowledge databases, such as RegulonDB (*E. coli*; Gama-Castro et al., 2008) and MIPS (yeast; Mewes et al., 2002). Functional congruence of clusters of coexpressed genes is a popular validation measure for clustering algorithms (Datta and Datta, 2006) and in principle can also be used for the validation of inferred networks.

3.3. NETWORK-BASED MEASURES

The third type of measures for assessing the performance of an inference algorithm considers the network structure explicitly.

[3]For example, the mutual information threshold I_0, used in the CLR algorithm.

That means, in contrast to the general statistical measures and also the ontology-based measures, network-based measures can only be used if there is a network that underlies the problem.

3.3.1. Global network-based measures

A measure that makes explicit use of the network structures of the true (G) and inferred ($\hat{G}$) network was proposed by Zhao et al. (2008). They suggested to use,

$$D(G,\hat{G}) = \frac{1}{|E|}\left(\sum_{A_{ij}=0,\ \hat{A}_{ij}=1} d_G(i,j) + \sum_{A_{ij}=1,\ \hat{A}_{ij}=0} 1\right), \qquad (32)$$

the weighted sum of false-positive edges (first term) plus the false-negative edges (second term). This measure is not only asymmetric in its arguments but also gives a different weight to type-1 (false positives) respectively type-2 (false negatives) errors. For type-1 errors, in the true network G the Dijkstra distance (Dijkstra, 1959) from node i to j, $d_G(i,j)$, is calculated whereas for type-2 errors each false-negative edge is assigned a constant weight of "1." Overall, $D(G,\hat{G})$ is a global measure, as were all measures discussed in section 1, which assesses the performance of an inference algorithm holistically.

3.3.2. Local network-based measures

In contrast to all above measures, which were *global* measures, we present finally *local* network-based measures introduced in Altay and Emmert-Streib (2010b), Emmert-Streib and Altay (2010). These local network-based measures are based on *ensemble data*, $\mathcal{D} = \{D_1(G),\ldots, D_E(G)\}$, and the availability of a reference network that represents the "true" regulatory network G. Ensemble data means that there is more than one data set available from the biological phenomenon under investigation. This ensemble could be obtained either by bootstrapping from one data set or from simulation studies. The apparent advantage of ensemble data is that they allow to quantify the variability of the population that underlies the data. After having obtained an ensemble of estimated networks $\mathcal{G}^e = \{G_i^e\}_{i=1}^E$ from $\mathcal{D} = \{D_1(G),\ldots, D_E(G)\}$ by application of an inference algorithm, we can obtain the true positive rates ($\{TPR_{ij}\}$) of each edge and the true negative rates ($\{TNR_{ij}\}$) for each non-edge with respect to the underlying true network structure. For example, we can estimate the TPR of an edge between gene i and j, which is present in G, by

$$TPR_{ij} = \frac{\#\text{ of edges present in } \mathcal{G}^e}{E} \qquad (33)$$

which corresponds to $Pr(\text{gene } i \text{ interacts with gene } j \,|\, \mathcal{G}^e)$. Analogously, we can estimate the TNR.

Principally, any combination of true positive and true negative rates would result in a valid network-based measure consisting, e.g., of network motifs, subnetworks, or even only of individual edges. For such a measure representing a structural region within a network it is then possible to estimate its reconstruction rate. To provide a concrete example, we give the reconstruction rate of a three-gene motif that is given by the chain in **Figure 2**

$$p = \frac{1}{3}\left(\mathrm{TPR}(\mathrm{X}\to\mathrm{Y}) + \mathrm{TPR}(\mathrm{Y}\to\mathrm{Z}) + \mathrm{TNR}(\mathrm{X}\not\to\mathrm{Z})\right). \qquad (34)$$

Also, if biological information about the genes in the network is available, this can also be used to define appropriate measures. For example, one could obtain a reconstruction rate for all interactions that are connected with transcription factors or of particular biological pathways involving only certain genes as defined, e.g., via the gene ontology database. Further examples of such measures can be found in Altay and Emmert-Streib (2010b), Emmert-Streib and Altay (2010).

4. COMPARISON OF INFERENCE METHODS

A question that is of practical relevance is which of the discussed inference methods, listed in **Figure 6**, is preferred. Unfortunately, there is no study that compared all these methods with each other. However, there are several studies that compared subsets thereof. For example, in Altay and Emmert-Streib (2010a, 2011) RN, ARACNE, CLR, MRNET, and C3NET have been compared for a large variety of different conditions, including different network structures, sample sizes, and noise levels. For all these conditions, ensemble data have been generated that allow the application of local network-based measures providing the most detailed information about the inference characteristics of a method. Overall, it has been found that C3NET and MRNET perform better than ARACNE, CLR, and RN (in this order). There are two main differences between C3NET and MRNET. First, the computational complexity of C3NET is $O(N^2)$ and for MRNET it is between $O(N^2)$ and $O(N^3)$ which is difficult to quantify exactly because of the iterative nature of the second step employed by MRNET. For the conducted simulations which involved networks in the order of $O(10^2)$ genes this was tractable, however, for real biological data the number of genes can reach over 10,000 causing series problems. In contrast, C3NET has been applied to a bootstrap

correlation-based methods

name	key method	software
coexpression	correlation	WGCNA
Asymmetric-N	correlation, ranking	no
SEM	genetic algorithm	no
ParCorA	partial correlation first order	ParCorA
GGM	full order partial correlation	GeneNet
BN	partial correlation n-th order	Hartemink

mutual information-based methods

name	key method	software
RN	mutual information	MINET
ARACNE	mutual information, DPI	MINET
CLR	mutual information, background	MINET
C3NET	maximum mutual information	C3NET
SA-CLR	mutual information, synergy	no
MRNET	mutual informations	MINET
MI3	three-way mutual information	MI3
CMI	conditional mutual information	no
MI-CMI,PCA-CMI	MI, conditional mutual information	Zhang
Liang & Wang	MI, conditional mutual information	no

FIGURE 6 | Overview of inference methods to reconstruct regulatory networks from expression data. First column, name of the method. Second column, methodological base of the method. Third column, name of the software package, if available. Horizontal classification, first panel, linear methods (correlation-based). Second panel, non-linear methods (information-based). Darker colors represent extended methods.

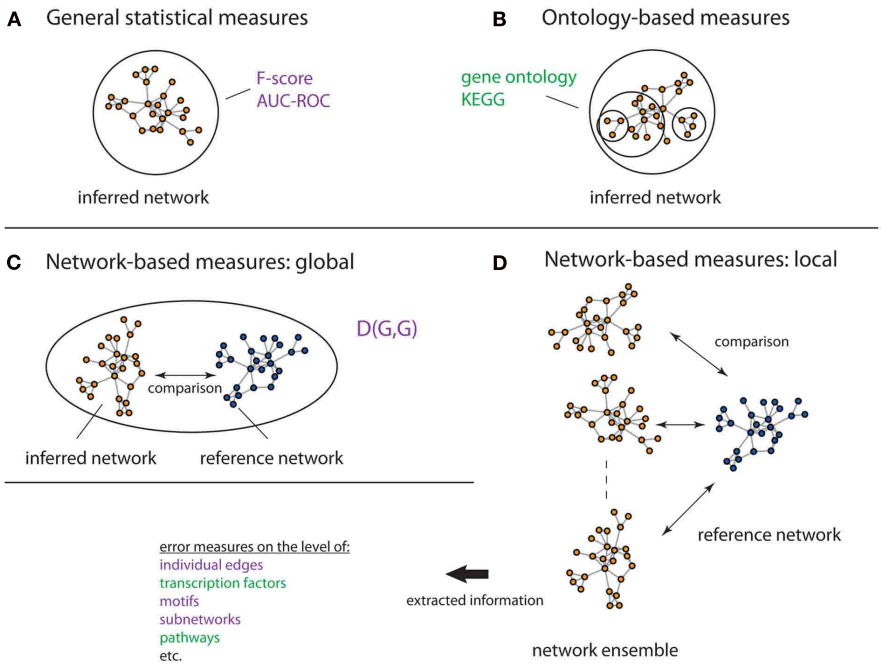

FIGURE 7 | Overview of different evaluation measures. Here, a circle/oval represents the entity that is assessed by a measure, which can be seen as its *resolution*. The figure visualizes general statistical measures **(A)**, ontology-based measures **(B)**, global network-based measures **(C)** and local network-based measures **(D)**. The lack of any circle/oval in **(D)** shall indicate that there is no restriction to any particular resolution level. The measures in green incorporate biological information, whereas the measures in purple do not.

ensemble from B cell lymphoma for 9.684 genes (de Matos Simoes et al., submitted). Second, due to the conservative character of C3NET, which is currently the most conservative method of all inference methods, the number of obtained interactions is easier to deal with than for other methods. For example, in Basso et al. (2005) ARACNE was applied to the same B cell data, inferring about 130,000 interactions whereas C3NET inferred about 10,000 interactions. Given the current impossibility to experimentally validate tens of thousands of interactions by wet lab experiments a conservative subnetwork having a low false-positive rate (Altay and Emmert-Streib, 2010a, 2011) is advantages.

We would like to remark that the methods studied in Altay and Emmert-Streib (2010a, 2011) were selected on the ground of their conceptual similarity to provide a fair comparison, but they have also a (relatively) low computational complexity which allows the investigation for ensemble data and not just single data sets. Many other methods listed in **Figure 6** have a higher computational complexity forcing a comparative analysis to be performed with small networks, typically involving $O(10^1)$ genes, and individual data sets rather than ensembles. For example, in Werhli et al. (2006) RN, GGM, and BN have been compared for a network consisting of only 11 genes because of the high computational complexity of a BN. As a result, it is reported that GGM and the BN outperform RN and perform similarly compared to each other.

5. DISCUSSION

From the discussions of the preceding methods for the structural inference of regulatory networks one could get the impression that their number is sheerly unlimited as well as the principle ideas behind them. In our opinion the first point is probably true, the latter not. In order to see this more clearly we would like to outline the general procedure underlying the development of each method. First, a principle idea or hypothesis is raised about a mechanism for the inference of regulatory networks and, second, a method is conceived that could accomplish this. Formally, the first part relates to the conceptual or qualitative framework of a method whereas the latter refers to its quantitative realization, e.g., in form of statistical estimators. In order to learn about two classic inference algorithms, embodying two different conceptual ideas, we presented the IC (inductive causation) algorithm (Pearl, 2000) and the MWST (maximum weight spanning tree) algorithm (Chow and Liu, 1968). For the IC algorithm the conceptual core is d-separation, and for the MWST this is the DPI. When considered from this perspective, the inference algorithms shown in **Figure 6** can be grouped as follows. ARACNE is based on the idea of the MWST, whereas all other methods except SEM, SA-CLR, and MI3 are either close or approximate adaptations of d-separation. Here approximate adaptation of d-separation means that an algorithm employs d-separation in a certain way, for instance up to a fixed order of the correlation coefficient or the mutual information (like ParCorA or MI-CMI), but not in the exhaustive way as described for the IC algorithm. From this perspective follow at least three implications. First, the number of conceptually different ideas is very limited if technical approximations are not counted as different. Second, when one of the above inference algorithms represents only an "approximation" of one

of the two classic conceptual frameworks its theoretical inference abilities would strictly speaking require a new mathematical proof because new assumptions made may not translate to these methods. Third, due to the fact that most methods presented in **Figure 6** lack such a strict mathematical proof our knowledge about their actual abilities is based on numerical studies. This does not mean that numerical studies are not capable of allowing a detailed analysis but that these investigations need to be conducted thoroughly acknowledging the statistical and biological nature of the problem. The former means that numerical studies need to be based on *ensemble data* in order to capture characteristics of the population, and the latter means that network-based metrics should be applied due to the heterogeneous inferability of local components of regulatory networks, as found in Altay and Emmert-Streib (2010b) and Emmert-Streib and Altay (2010).

We would like to finish this review with a brief outlook on future directions. In the introduction, we mentioned that due to the nature of gene expression data, which do not allow to derive unique predictions about the underlying molecular interactions between gene products, the resulting gene regulatory networks represent a mixture of a transcriptional regulatory network and a protein interaction network. For example, in Altay and Emmert-Streib (2010a) gene expression data from *E. coli* have been analyzed by inferring a regulatory network. Among the verified interactions found from comparing estimated interactions with experimentally reported results from the literature, transcriptional regulatory interactions, and also protein-protein interactions have been found. Similar results for *S. cerevisiae* can be found in de Matos Simoes and Emmert-Streib (submitted). In this study, over 800 protein-protein interactions from BioGRID (Breitkreutz et al., 2008) have been identified in the inferred network. For a general discussion about the connection between gene expression and protein interactions, see Grigoriev (2001), Jansen et al. (2002).

In order to obtain more refined predictions about the type of interactions and also to improve the inference performance of the methods, complementary information from other types of high-throughput data is needed. For example, data from ChIP-chip or ChIP-Seq experiments could be used to obtain information about the potential gene targets of transcription factors, similarly, proteomics data could be employed to reveal protein-protein interactions. Ideally, information from all three data types (ChIP-Seq, gene expression, and proteomics) should be integrated to infer a more detailed network with a clearer interpretation of the inferred interactions between the gene products. Sporadically, methods have been already pioneered for such an integration (Nariai et al., 2005; Xing and van der Laan, 2005). However, due to the increased experimental effort and its associated costs the combined availability of several different types of large-scale high-throughput data sets, as necessary for such an analysis, is still rare. This lack of data hampered so far the systematic development of integrative methods. However, with the increasing use of next-generation sequencing technologies this might change is the near future. Other data integration approaches that need to be addressed are the combination of, e.g., gene expression data, from different experiments or platforms (Belcastro et al., 2011). Due to the need for a normalization of such expression data, a pooling of these data is usually prohibitive making an analysis difficult.

Another very important topic, aside data integrative methods, relates to the generation of the data themselves. Specifically, in this review we focused on observational data only, however, experimental data consisting of gene interventions or perturbations form a very fruitful source of information that could be systematically exploited (Fröhlich et al., 2008; Markowetz, 2010; Pinna et al., 2010; Yip et al., 2010).

This discussion emphasizes the need for a clear conceptual distinction between different methods and the information they are based on.

6. CONCLUSION

In this paper we presented a systematic overview of methods for inferring gene regulatory networks. Although this field is currently vastly expanding making it very difficult to obtain such an overview, we assumed two different perspectives that allowed to categorize inference algorithms sensibly. The first perspective was based on the dynamical assumptions (linear vs. non-linear) methods make about the underlying data. The second considered the methods through the lense of classic approaches which use either d-separation or the DPI (Chow and Liu, 1968; Verma and Pearl, 1991; Spirtes et al., 1993). We want to conclude this article by mentioning that the inference of regulatory networks may not only help in gaining a better understanding of the *normal* physiology of a cell, but also in elucidating the molecular basis of diseases (Emmert-Streib, 2007; Chen et al., 2008b; Jiang et al., 2008).

ACKNOWLEDGMENTS

Frank Emmert-Streib would like to thank Simon Tavaré and Florian Markowetz for fruitful discussions and the Department for Employment and Learning through its "Strengthening the all-Island Research Base" initiative and the School of Medicine, Dentistry and Biomedical Sciences for financial support. Gökmen Altay is funded by Cancer Research UK.

REFERENCES

Agrawal, H. (2002). Extreme self-organization in networks constructed from gene expression data. *Phys. Rev. Lett.* 89, 268702.

Altay, G., and Emmert-Streib, F. (2010a). Inferring the conservative causal core of gene regulatory networks. *BMC Syst. Biol.* 4, 132. doi:10.1186/1752-0509-4-132

Altay, G., and Emmert-Streib, F. (2010b). Revealing differences in gene network inference algorithms on the network-level by ensemble methods. *Bioinformatics* 26, 1738–1744.

Altay, G., and Emmert-Streib, F. (2011). Structural Influence of gene networks on their inference: analysis of C3NET. *Biol. Direct* 6, 31.

Anastassiou, D. (2007). Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.* 3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17299419

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3, 78.

Barabási, A. L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human b cells. *Nat. Genet.* 37, 382–390.

Beirlant, J., Dudewicz, E., Gyorfi, L., and van der Meulen, E. (1997). Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.* 6, 17–39.

Belcastro, V., Siciliano, V., Gregoretti, F., Mithbaokar, P., Dharmalingam, G., Berlingieri, S., Iorio, F., Oliva, G., Polishchuck, R., Brunetti-Pierri, N., and di Bernardo, D. (2011). Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. *Nucleic Acids Res.* 39, 8677–8688.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188.

Bollen, K. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Bonferroni, E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze* 8, 3–62.

Breitkreutz, B.-J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V., Dolinski, K., and Tyers, M. (2008). The BioGRID interaction database: 2008 update. *Nucleic Acids Res.* 36(Suppl. 1), D637–D640.

Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research.* New York: The Guilford Press.

Butte, A., and Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomp.* 5, 415–426.

Carter, S. L., Brechbuhler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 20, 2242–2250.

Cerulo, L., Elkan, C., and Ceccarelli, M. (2010). Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinformatics* 11, 228. doi:10.1186/1471-2105-11-228

Chen, G., Larsen, P., Almasri, E., and Dai, Y. (2008a). Rank-based edge reconstruction for scale-free genetic regulatory networks. *BMC Bioinformatics* 9, 75. doi:10.1186/1471-2105-9-75

Chen, Y., Zhu, J., Lum, P., Yang, X., Pinto, S., MacNeil, D. J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S. K., Leonardson, A., Castellini, L. W., Wang, S., Champy, M. F., Zhang, B., Emilsson, V., Doss, S., Ghazalpour, A., Horvath, S., Drake, T. A., Lusis, A. J., and Schadt, E. E. (2008b). Variations in DNA elucidate molecular networks that cause disease. *Nature* 452, 429–435.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. *Networks* 112, 121–130.

Chickering, D., Geiger, D., and Heckerman, D. (1995). "Learning bayesian networks: search methods and experimental results," in *Proceedings of Fifth Conference on Artificial Intelligence and Statistics* (Society for Artificial Intelligence in Statistics), 569–595.

Chow, C., and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory* 14, 462–467.

Cover, T., and Thomas, J. (1991). *Information Theory*. Hoboken, NJ: John Wiley & Sons, Inc.

Crick, F. (1970). Central dogma of molecular biology. *Nature* 227, 561–563.

Datta, S., and Datta, S. (2006). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics* 7, 397. doi:10.1186/1471-2105-7-397

Daub, C., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions – an improved similarity measure for analysing gene expression data. *BMC Bioinformatics* 5, 118. doi:10.1186/1471-2105-5-118

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67–103.

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20, 3565–3574.

de Matos Simoes, R., and Emmert-Streib, F. (2011). Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks. *PLoS ONE* 6, e29279. doi:10.1371/journal.pone.0029279

de Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8, 717–729.

Dean, T., and Kanazawa, K. (1990). A model for reasoning about persistence and causation. *Comput. Intell.* 5, 142–150.

Dempster, A. (1972). Covariance selection. *Biometrics* 28, 157–175.

Dijkstra, E. (1959). A note on two problems in connection with graphs. *Numer. Math.* 1, 269–271.

Dudoit, S., Gilbert, H., and van der Laan, M. (2008). Resampling-based empirical bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: focus on the false discovery rate and simulation study. *Biom. J.* 50, 716–744.

Dudoit, S., and van der Laan, M. (2007). *Multiple Testing Procedures with Applications to Genomics.* New York: Springer.

Emmert-Streib, F. (2007). The chronic fatigue syndrome: a comparative pathway analysis. *J. Comput. Biol.* 14, 961–972.

Emmert-Streib, F., and Altay, G. (2010). Local network-based measures to assess the inferability of different regulatory networks. *IET Syst. Biol.* 4, 277–288.

Emmert-Streib, F., and Glazko, G. (2011). Network biology: a direct approach to study biological function. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 3, 379–391.

Ernst, J., Beg, Q. K., Kay, K. A., Balazsi, G., Oltvai, Z. N., and Bar-Joseph, Z. (2008). A semi-supervised method for predicting transcription factor-gene interactions in *Escherichia coli*. *PLoS Comput. Biol.* 4, e1000044. doi:10.1371/journal.pcbi.1000044

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5, 8. doi:10.1371/journal.pbio.0050008

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian network to analyze expression data. *J. Comput. Biol.* 7, 601–620.

Friedman, N., Nachman, I., and Pe'er, D. (1999). "Learning Bayesian network structure from massive datasets: the 'Sparse Candidate' algorithm," in *Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, (Society for Artificial Intelligence in Statistics), 206–215.

Fröhlich, H., Beißbarth, T., Tresch, A., Kostka, D., Jacob, J., Spang, R., and Markowetz, F. (2008). Analyzing gene perturbation screens with nested effects models in r and bioconductor. *Bioinformatics* 24, 2549–2550.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A. M., Treviño-Quintanilla, L., and Collado-Vides, J. (2008). RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* 36(Suppl. 1), D120–D124.

Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 29, 3513–3519.

Hache, H., Lehrach, H., and Herwig, R. (2009). Reverse engineering of gene regulatory networks: a comparative study. *EUROSIP J. Bioinform. Syst. Biol.* Available at: http://www.hindawi.com/journals/bsb/2009/617281/

Hartemink, A., Gifford, D., Jaakkola, T., and Young, R. (2001). "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks," in *Pacific Symposium on Biocomputing* (NJ: World Scientific), 422–433.

Horvath, S., and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* 4, e1000117. doi:10.1371/journal.pcbi.1000117

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics* 19, 2271–2282.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res.* 12, 37–46.

Jiang, W., Li, X., Rao, S., Wang, L., Du, L., Li, C., Wu, C., Wang, H., Wang, Y., and Yang, B. (2008). Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. *BMC Syst. Biol.* 2, 72. doi:10.1186/1752-0509-2-72

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrica* 34, 183–202.

Jöreskog, K. (1973). "A general method for estimating a linear structural equation system," in *Structural Equation Models in the Social Sciences*, eds A. Goldberger and O. Duncan (New York: Seminar Press), 85–112.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Khan, S., Bandyopadhyay, S., Ganguly, A., Saigal, S., Erickson, D., Protopopescu, V., and Ostrouchov, G. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 76, 026209.

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E* 69, 066138.

Lehmann, E., and Romano, J. (2005). Generalizations of the familywise error rate. *Ann. Stat.* 33, 1138–1154.

Li, H., and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* 7, 302–317.

Liang, K., and Wang, X. (2008). Gene regulatory network reconstruction using conditional mutual information. *EURASIP J. Bioinform. Syst. Biol.* 2008, 253894.

Liu, W., Lähdesmäki, H., Dougherty, E., and Shmulevich, I. (2008). Inference of Boolean networks using sensitivity regularization. *EURASIP J. Bioinform. Syst. Biol.* 2008, 780541.

Luo, W., Hankenson, K. D., and Woolf, P. J. (2008). Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC Bioinformatics* 9, 467. doi: 10.1186/1471-2105-9-467

Magwene, P. M., and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.* 5, R100.

Margolin, A., and Califano, A. (2007). Theory and limitations of genetic network inference from microarray data. *Ann. N. Y. Acad. Sci.* 1115, 51–72.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7, S7. doi:10.1186/1471-2105-7-S1-S7

Markowetz, F. (2010). How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS Comput. Biol.* 6, e1000655. doi:10.1371/journal.pcbi.1000655

Markowetz, F., and Spang, R. (2007). Inferring cellular networks–a review. *BMC Bioinformatics* 8, S5. doi:10.1186/1471-2105-8-S6-S5

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika* 19, 97–116.

Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34.

Meyer, P. E., Kontos, K., Latiffe, F., and Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EUROSIP J. Bioinform. Syst. Biol.* 2007, Article ID 79879.

Mordelet, F., and Vert, J.-P. (2008). SIRENE: supervised inference of regulatory networks. *Bioinformatics* 24, i76–i82.

Nariai, N., Tamada, Y., Imoto, S., and Miyano, S. (2005). Estimating gene regulatory networks and protein-protein interactions of *Saccharomyces cerevisiae* from multiple genome-wide data. *Bioinformatics* 21(Suppl. 2), 206–212.

Olsen, C., Meyer, P., and Bontempi, G. (2009). On the impact of entropy estimator in transcriptional regulatory network inference. *EURASIP J. Bioinform. Syst. Biol.* 2009, 308959.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan-Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Penfold, C. A., and Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface Focus* 1, 857–870.

Perrin, B.-E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d'Alche Buc, F. (2003). Gene networks inference using dynamic Bayesian networks. *Bioinformatics* 19, ii138–ii148.

Pinna, A., Soranzo, N., and de la Fuente, A. (2010). From knockouts to networks: establishing direct cause-effect relationships through graph analysis. *PLoS ONE* 5, e12912. doi:10.1371/journal.pone.0012912

Rosenbaum, P. (2002). *Observational Studies*. New York: Springer.

Rubin, R. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701.

Schadt, E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223.

Schäfer, J., and Strimmer, K. (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21, 754–764.

Shipley, B. (2000). *Cause and Correlation in Biology*. Cambridge, NY: Cambridge University Press.

Soranzo, N., Bianconi, G., and Altafini, C. (2007). Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics* 23, 1640–1647.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. New York: Springer.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 18, S231–S240.

Stolovitzky, G., and Califano, A. (eds). (2007). *Reverse Engineering Biological Networks: Opportunities and Challenges in Computational Methods for Pathway Inference*. Boston: Wiley-Blackwell.

Stolovitzky, G., Prill, R., and Califano, A. (2009). Lessons from the DREAM 2 challenges. *Ann. N. Y. Acad. Sci.* 1158, 159–195.

Storey, J., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445.

Verma, T., and Pearl, J. (1990). "Causal networks: semantics and expressiveness," in *Proceedings of the 4th Workshop on Uncertainly in Artificial Intelligence*, Mountain View, CA, 352–359.

Verma, T., and Pearl, J. (1991). "Equivalence and synthesis of causal models," in *Proceedings of the 6th Workshop on Uncertainty in Artificial Intelligence*, Cambridge, MA, 220–227.

Watkinson, J., Liang, K., Wang, X., Zheng, T., and Anastassiou, D. (2009). Inference of regulatory gene interactions from expression data using three-way mutual information. *Ann. N. Y. Acad. Sci.* 1158, 302–313.

Werhli, A., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics* 22, 2523–2531.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: John Wiley & Sons.

Wille, A., and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.* 4, 32.

Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Buhlmann, P. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.* 5, R92.

Wolfe, C., Kohane, I., and Butte, A. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics* 6, 227. doi:10.1186/1471-2105-6-227

Wright, S. (1921). Correlation of causation. *J. Agric. Res.* 20, 557–585.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Stat.* 5, 161–215.

Xing, B., and van der Laan, M. (2005). A causal inference approach for constructing transcriptional regulatory networks. *Bioinformatics* 21, 4007–4013.

Xiong, M., Li, J., and Fang, X. (2004). Identification of genetic networks. *Genetics* 166, 1037–1052.

Yip, K. Y., Alexander, R. P., Yan, K.-K., and Gerstein, M. (2010). Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS ONE* 5, e8121. doi:10.1371/journal.pone.0008121

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, 17.

Zhang, X., Zhao, X.-M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.-K., Liu, Z.-P., and Chen, L. (2011). Inferring gene regulatory networks from gene expression data by PC-algorithm based on conditional mutual information. *Bioinformatics*. http://www.ncbi.nlm.nih.gov/pubmed/22088843

Zhao, W., Serpendin, E., and Dougherty, E. R. (2008). Inferring connectivity of genetic regulatory networks using information-theoretic criteria. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 5, 262–274.

Zhou, X., Kao, M., and Wong, W. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U.S.A.* 99, 12783–12788.

Zou, M., and Conzen, S. D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* 21, 71–79.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 December 2011; accepted: 10 January 2012; published online: 03 February 2012.

Citation: Emmert-Streib F, Glazko GV, Altay G and de Matos Simoes R (2012) Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Front. Gene. ***3****:8. doi: 10.3389/fgene.2012.00008*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.