BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 20 2013, pages 2603–2609
doi:10.1093/bioinformatics/btt438

Gene expression

Advance Access publication August 16, 2013

A Turing test for artificial expression data

Robert Maier, Ralf Zimmer and Robert Küffner*

Department of Informatics, Ludwig-Maximilians Universität, 80333 München, Germany

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: The lack of reliable, comprehensive gold standards complicates the development of many bioinformatics tools, particularly for the analysis of expression data and biological networks. Simulation approaches can provide provisional gold standards, such as regulatory networks, for the assessment of network inference methods. However, this just defers the problem, as it is difficult to assess how closely simulators emulate the properties of real data.

Results: In analogy to Turing's test discriminating humans and computers based on responses to questions, we systematically compare real and artificial systems based on their gene expression output. Different expression data analysis techniques such as clustering are applied to both types of datasets. We define and extract distributions of properties from the results, for instance, distributions of cluster quality measures or transcription factor activity patterns. Distributions of properties are represented as histograms to enable the comparison of artificial and real datasets. We examine three frequently used simulators that generate expression data from parameterized regulatory networks. We identify features distinguishing real from artificial datasets that suggest how simulators could be adapted to better emulate real datasets and, thus, become more suitable for the evaluation of data analysis tools.

Availability: See http://www2.bio.ifi.lmu.de/~kueffner/attfad/ and the supplement for precomputed analyses; other compendia can be analyzed via the CRAN package attfad. The full datasets can be obtained from http://www2.bio.ifi.lmu.de/~kueffner/attfad/data.tar.gz.

Contact: robert.kueffner@bio.ifi.lmu.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 21, 2012; revised on June 28, 2013; accepted on July 26, 2013

1 INTRODUCTION

The analysis of expression data involves a range of bioinformatics tools. They are used for outlier removal and normalization (Fundel *et al.*, 2008), for clustering (Eisen *et al.*, 1998), for finding relevant biological processes (Draghici *et al.*, 2003; Naeem *et al.*, 2012) as well as for the inference and modeling of gene regulatory interactions (GRIs) (Küffner *et al.*, 2010, 2012). The improvement of such tools frequently depends on insights into properties of transcriptional regulation. To better understand such properties and to facilitate further tool development, several simulators have been developed for various areas of expression analysis.

Another reason for conducting simulation studies is to provide gold standards that are not available for real data. In case of clustering, the number of gene clusters and the cluster–gene associations are usually not known for real data but can be defined easily within an *in silico* framework (Wu *et al.*, 2004; Yeung *et al.*, 2001). For the assessment of tools, it is important that simulators generate realistic expression patterns.

Similar techniques are applied to analyze whether gene sets associated with biological processes exhibit a significant enrichment of differential expression. The assessment of enrichment tests on real data is difficult, as the genes or processes that exhibit differential activity between experimental conditions are usually not known before the actual experiment (Naeem *et al.*, 2012). As a remedy, enrichment tests can be assessed based on simulations of gene sets and their differential expression patterns (Ackermann and Strimmer, 2009; Efron and Tibshirani, 2007; Nam and Kim, 2008).

Furthermore, Albers *et al.* (2006) generated realistic data and noise distributions to examine data normalization approaches for two-dye arrays. Simulators are also used to examine specific theoretical assumptions [e.g. expression data distributions (Parrish *et al.*, 2009), single-cell data (Hebenstreit and Teichmann, 2011) and theoretical properties of GRIs (Wu and Chan, 2012)].

Simulators devised for studying GRI inference approaches are again motivated by the need for complete gold standard networks. Unlike the specific simulators, network simulators claim to more generally satisfy the mentioned requirements. They imitate experimentally determined GRIs and their dynamical properties and emulate other properties of real data such as data and noise distributions (Haynes and Brent, 2009; Pinna *et al.*, 2011; Schaffter *et al.*, 2011; Van den Bulcke *et al.*, 2006).

Network simulators have been applied successfully to assess algorithms that reconstruct GRIs from expression data, e.g. via community-wide blind assessments (Marbach *et al.*, 2012; Pinna *et al.*, 2011). Therefore, simulators implement elaborate experimental settings, for instance, by dealing with biological and technical replicates and by generating data for gene knockouts, various perturbations and time series data (Schaffter *et al.*, 2011). Some simulators also take into account sequence variations in promoter regions and coding sequences (Pinna *et al.*, 2011). Instead of random topologies, many simulators use realistic topologies, such as scale-free networks (Haynes and Brent, 2009) or networks of experimentally determined GRIs (Schaffter *et al.*, 2011; Van den Bulcke *et al.*, 2006).

Algorithms can be benchmarked using simulators in a meaningful way if generated, and real data are sufficiently similar (Haynes and Brent, 2009; Pinna *et al.*, 2011; Schaffter *et al.*, 2011; Van den Bulcke *et al.*, 2006). As this is difficult to show, *in silico* models are sometimes considered insufficient evidence for method enhancements. This is owing to our lack of a good understanding of the distribution of expression levels across the

*To whom correspondence should be addressed.

various biological states and the correlation between biological replicates (Rocke *et al.*, 2009).

The present study evaluates to which extent current state-of-the-art simulators are able to generate realistic datasets and gold standards. We propose characteristic histograms to compare artificial and real expression compendia. The histograms are derived by applying a range of expression data analysis approaches. We provide data on our Web site and *R* software to apply the suggested analyses on the compendia examined in this article or defined by the user. The properties of real and artificial datasets thus become amenable to a mutual pairwise comparison. For the generation of artificial datasets, we focus on network simulators, as they are most frequently used in practice and most general in their scope.

2 DATASETS

To compare real and artificial expression compendia, we obtained a range of datasets. This section briefly describes the real datasets and experimentally derived networks (see the cited original publications for more information). Artificial datasets and networks are described in the Methods section.

2.1 Expression data

Four microarray compendia with several hundred microarray measurements were obtained for *Escherichia coli* and *Saccharomyces cerevisiae* from the M3D Database (Faith *et al.*, 2008) (denoted with index 1) and from the DREAM5 competition (Marbach *et al.*, 2012) (denoted with index 2). Measurements as well as annotations derived from these sources are already rendered comparable across different experiments and are thus suited to automated analyses. Besides wild-type measurements, experimental conditions represent (combinations of) drug, environmental and gene perturbations. Some of the drug or environmental perturbations are provided as time course measurements. We considered each time point as a separate condition. Each condition may contain multiple replicates. Before all analyses, we $\log_2$-transformed the measurements. Each dataset is represented as an expression matrix M, where rows correspond to the genes and columns to the measurements, i.e. the element $m_{i,j}$ of this matrix is the jth measurement of the ith gene.

2.2 Gene regulatory networks

The topology of *E.coli* gene regulatory networks was taken from RegulonDB (Gama-Castro *et al.*, 2011). RegulonDB is a database of gene regulatory relationships that are both experimentally validated and manually curated. The *S.cerevisiae* network was derived by large-scale Chromatin Immuno-Precipitation (ChIP) assays post-processed by MacIsaac *et al.* (2006). These large-scale networks cover only transcription factor–target gene (TF:TG) interactions and ignore other kinds of interactions, e.g. based on microRNA–target interactions.

3 METHODS

3.1 Simulators and their areas of application

GeneNetWeaver (GNW) (Schaffter *et al.*, 2011) generates mRNA readouts by simulation of transcription and translation via stochastic differential equations that are randomly parameterized to construct executable models. Additionally, measurement noise is added to the gene expression data after simulation. Regulation takes place at the transcriptional level only, i.e. other layers of regulation such as protein–protein interactions, metabolites or cellular states are not considered. Instead of random graphs, GNW uses the known biological GRI networks, e.g. from *E.coli* and *S.cerevisiae*, but removes auto-regulatory loops. GNW data do not vary over several orders of magnitude and, thus, correspond to log-transformed real data.

Three levels of experiments are supported where the variability of simulated expression runs increases from level one to three. Technical replicates as the first level are based on the same model and parameterization; the generated data are distinguished by noise only. The second level models biological replicates by slight random increases or decreases of the basal levels of all genes. Finally, distinct conditions are applied to the simulation mimicking time course, (multi or single) gene deletion or gene overexpression experiments. This dataset was specifically designed to mimic the types of experiments performed in the DREAM5 *E.coli* dataset (Marbach *et al.*, 2012), which is comprehensively described (http://wiki.c2b2.columbia.edu/dream/index.php?title=D5c4). The GNW dataset consists of 1643 genes, 195 of which are transcription factors, and 805 microarray experiments.

Gene REgulatory Network Decoding Evaluations tooL (GRENDEL; Haynes and Brent, 2009), in contrast to GNW, does not directly derive the topology of the simulated gene regulatory networks from real networks. Instead, a topology is modeled that reproduces in- and out-degree distributions of real networks.

Each simulated gene is paired with an *S.cerevisiae* gene to use experimental data, such as translation rate, protein decay rate, mRNA decay rate and mRNA transcription rate for the parameterization of a deterministic ordinary differential equation (ODE) simulation. However, topology modeling does not guarantee that paired genes in the yeast network are similarly close in the modeled network. Owing to the lack of suitable data, parameters determining the dynamical properties of GRIs are chosen randomly. Genes may not only be regulated by transcription factors, but also by environmental signals that can have instantaneous effects on other genes. Measurement noise is added to the data according to a log-normal distribution. Thus, only technical replicates and experimental conditions are modeled, and biological replicates are not taken into account.

Expression data were simulated for 100 genes across 300 measurements. This population was modeled by varying the efficiency of transcription and mRNA degradation across samples independently for each gene. The topology of the network is described by an exponential distribution for in-degrees and a power-law distribution for out-degrees and was fit to reflect the degree distributions of the transcriptional network in *S.cerevisiae*. GRENDEL uses the Systems Biology Markup Language ODE Solver library to deterministically integrate the ODEs that define the dynamical system. To the resulting expression data, simulated experimental noise is then added according to a log-normal distribution.

SynTReN. The network topology of a simulation with SynTReN (Van den Bulcke *et al.*, 2006) is constructed from subsets of the known gene regulatory networks from *E.coli* and *S.cerevisiae*, which is similar to GNW's approach. The differential equations used for the simulation are based on Michaelis–Menten and Hill kinetic equations and allow us to simulate independent as well as synergistic interactions. Although pre-generated datasets were used in case of GNW (Marbach *et al.*, 2012) and GRENDEL (Haynes and Brent, 2009), we modeled a network of 300 genes with SynTReN and generated an expression compendium of 200 measurements using default parameters.

3.2 General approach to dataset assessment

We suggest a procedure for the extraction of individual dataset properties and their comparison across datasets that consists of four steps. We

(1) apply standard expression data analysis techniques,

(2) analyze the results to generate specific distributions,

(3) visualize the distributions by histograms and, finally,

(4) compare different datasets based on histogram overlaps.

Steps 1 and 2 depend on the details of the individual analyses (see next section), while steps 3 and 4 are the same across analyses.

In step 3, we plot histograms from analysis-specific distributions. For each type of histogram, we manually picked the smallest number of histogram bins (between 20 and 100, this can be changed in the R-package) that appropriately displayed histograms for a property. Histograms are computed by equally distributing the bins between the lowest and highest values of the compared histograms. Unless noted otherwise, we plot normalized densities in the histograms where the bars receive an average density of 1. Although graphical representations enable the closer examination of histograms, the largest differences are identified via overlap scores.

For each property (step 4), we calculate overlaps to compare pairs of histograms a and b by dividing the minimum density by the maximum density over n bins: $overlap = \sum_{i=1}^{n} min(a_i, b_i) / \sum_{i=1}^{n} max(a_i, b_i)$.

3.3 Characterization of dataset properties

For the comparison of datasets, we characterize various properties by specific histograms derived from quality measures and standard gene expression analyses (Fig. 1).

Replicate noise histogram. Pearson's correlation coefficient ρ is computed for all pairs of replicate measurements contained in a given expression compendium. For example, if a measurement is repeated four times, six pairwise correlations are computed. Correlations of all pairs of replicated measurements are combined into one histogram visualizing pairs of replicates which exhibit given intervals of ρ.

Intensity histogram. For this analysis, all intensity measurements from a given expression compendium are combined into one histogram visualizing measurements with particular expression intensities. Such histograms are important to evaluate how effective approaches for data normalization are.

Range of gene expression. The expression range of individual genes is estimated from the difference between its 5% and 95% intensity quantiles observed in a given expression compendium. A corresponding histogram is created to show the histogram of gene ranges. Although a fold change of two is considered a substantial change in case of the real data, no such intuition exists in case of the artificial data that are scaled very differently. Therefore, we multiply the artificial ranges by a factor so that their median corresponds to the median of the compared real range histogram.

Silhouette coefficient (two histograms). To evaluate cluster quality, we use hierarchical average linkage clustering (Evisen *et al.*, 1998), using Pearson's ρ as the similarity measure. Clustering is performed separately for genes (using the corresponding expression matrix M) and microarrays (based on the transposed matrix M^T). The quality of the clustering at each of the successive node (i.e. gene or microarray) join steps is evaluated by Silhouette coefficients (Rousseeuw, 1987) that compare the average within-cluster distances with the average between-cluster distance. Thus, this coefficient evaluates the tightness and separation of clusters. For a dendrogram consisting of n nodes, n – 1 Silhouette values are obtained. Hierarchical clustering illustrates properties of co-regulation in expression data, i.e. the emergence of gene expression patterns across different experimental conditions.

TF and TG correlation (two histograms). To examine the correlation between TFs and regulated TGs, we generate three histograms. Correlation coefficients ρ are calculated for (i) all pairs of genes, (ii) all pairs of a TF and an annotated target as well as (iii) all pairs of targets that are regulated by the same set of TFs. Histogram (i) thus represents the background distribution of correlations between arbitrary pairs of genes. Finally, two difference diagrams are generated to show the enrichment of TF–TG correlations as the difference between (ii) and (i) and the enrichment of TG–TG correlations as the difference between (iii) and (i). The resulting difference diagrams depict to what extent highly correlated pairs are enriched (positive values) or depleted (negative values) for true TF–TG or TG–TG relationships. TF–TG correlations thus enable us to evaluate whether statistical dependencies exist between the expression levels of TFs and their targets (Marbach *et al.*, 2012; Wu and Chan, 2012). Alternatively, TG–TG correlations estimate the degree of co-regulation induced in pairs of genes regulated by the same TFs.

TF activity histogram. We determine the activity of TFs based on the expression profiles of their sets of TGs in the available GRI gold standard. Wilcoxon's nonparametric rank-sum method (WR test) (Mann and Whitney, 1947) is applied to test whether TF targets exhibit significant rank differences in comparison with other (non-targets) genes (Naeem *et al.*, 2012). The ranks are derived by sorting the genes based on their z-transformed absolute expression, i.e. for each condition i, the expression level $l_i(g)$ of gene g is transformed into a z-score $z_i(g)$ via $z_i(g) = (l_i(g) - m(g))/s(g)$ where m(g) and s(g) denote the mean and standard deviation of the expression levels of g, respectively. After

Section	Type of analysis	Motivation and questions	Figure: Simulator
4.1	Replicate noise distribution	Reproducibility of measurements, likelihood of outliers	2a: GNW
4.1	Intensity distribution	Choose normalization method, evaluate effectiveness	2b: GNW
4.1	Range of gene expression	Count distinguishable gene levels	2cd: GNW
4.2	Gene silhouette coefficient	Assess gene cluster quality	3a: Grendel
4.2	Chip silhouette coefficient	Assess microarray cluster quality	3b: Grendel
4.3	TF-TG correlation	Dependencies between expression of TFs and TGs	4a: SynTReN
4.3	TG-TG correlation	Co-regulation of TGs induced by common TFs	4b: SynTReN
4.3	TF activity distribution	Distribution of TF activities across conditions	4c: SynTReN

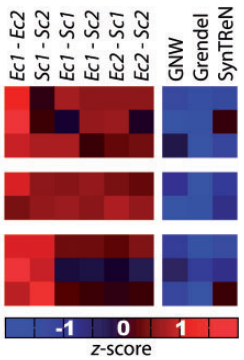

Fig. 1. Overview of tested properties and corresponding areas of expression data analysis. Expression analyses were performed to compare four real and three *in silico* expression compendia. For each type of analysis, figures in this article use *in silico* data of a different simulator. Additional analyses were performed across all simulators (see Web site and R software for more figures), and corresponding results are shown in the heatmap on the right. In the heatmap, different analyses (rows) compare various pairs of datasets (columns; E.c. = *E.coli*; S.c. = *S.cerevisiae*) based on histogram overlaps. The heatmap depicts the extent to which dataset properties are shared across compendia. Overlaps are shown in units of row-wise standard deviation (see online, red = high, blue = low overlap). In case of the three simulators, the average overlap to the four real expression compendia is shown. See section 4 for more details

this transformation, particularly high or low levels of a given gene can be clearly distinguished from its average levels. The WR test results in P values that we correct via Benjamini–Hochberg's (Benjamini and Yekutieli, 2001) procedure for multiple testing. Regulators are called active if they receive a corrected $P<0.05$ (Naeem *et al.*, 2012). For a given expression compendium, one histogram is constructed to show how many TFs are found active in particular subsets of the measurements.

4 RESULTS

We compare various properties between real and artificial datasets and find strong deviations (Fig. 1, heatmap). Properties and their deviations are discussed below in detail. Each subsection presents one type of property histogram and is exemplified in detail based on the data from one of the three simulators (Fig. 1). Results on the other simulators are briefly summarized at the end of each section (see availability for additional figures).

4.1 Simple data histograms (GNW)

Replicate noise histogram. We first examined the reproducibility of measurements via Pearson's correlation coefficient ρ (Fig. 2a). In case of artificial data, variations between replicates are due to the noise that simulators add to the generated expression data. In comparison with the real data where most replicates exhibit a very high ρ, replicate measurements in artificial data are substantially less correlated, which is due to excess amounts of added noise. In addition, real data usually include a number of outliers that show substantially lower correlations than the main distribution of replicate correlations. In contrast, no outliers were observed in case of artificial data where the peaks of the distributions were sharply bounded (see scatter plots as inset panels of Fig. 2a).

Intensity histogram. Although measured intensity histograms generally depend on the used microarray platform (arrays used in this study were performed on Affymetrix platforms), intensities usually exhibit positive values that are approximately normally distributed (Fig. 2b). In contrast, the simulated data histograms we examined were not normally distributed and peaked at an absolute expression level of 0. In addition, histograms derived from individual real microarrays show a much greater diversity as shown in Figure 2b.

Range of gene expression. In contrast to the overall histogram of expression intensities (Fig. 2b), Figure 2c visualizes the range of fold changes of genes. For instance, a value of 3 denotes the genes that exhibit fold changes of up to 8 between their 5% and 95% quantiles. While the distributions were sharply bounded in case of artificial data (with a peak at a fold range of 2), the real data exhibited a distribution with a long tail of genes covering a much broader spectrum of fold ranges.

Figure 2c further enables the estimation of the number of discrete expression levels or activity states assumed by individual genes. This is important for modeling approaches such as Bayesian (Needham *et al.*, 2006) or fuzzy logic models (Küffner *et al.*, 2010) requiring discretized values. A $\log_2$ fold range of 2 will correspond to three distinguishable gene levels if discrete sets are separated by a $\log_2$ fold change of 1. As demonstrated by Küffner *et al.* (2010), a gene model with three states is indeed sufficient to model all genes in the GNW dataset. However, in case of real data, the majority of genes could be suitably modeled with only two states (corresponding to the peak at 1 in Fig. 2c), while many genes require a larger number of states. A substantial number of genes showed fold ranges of >4.5 (132 of 4511 genes in *E.coli*) that accordingly exhibit ≥6 states.

Deviational patterns described here for GNW also apply to GRENDEL. On the other hand, a closer examination of the histograms (see availability for corresponding plots) reveals that data generated by SynTReN actually resemble real data histograms in case of the intensity and gene range properties. However, SynTReN's replicate noise distribution is extremely broad and thus resembles neither the real data nor the data generated by other simulators.

4.2 Cluster analyses (GRENDEL)

Silhouette coefficients. Functionally related genes show tight co-regulation patterns in real datasets that are clearly separated

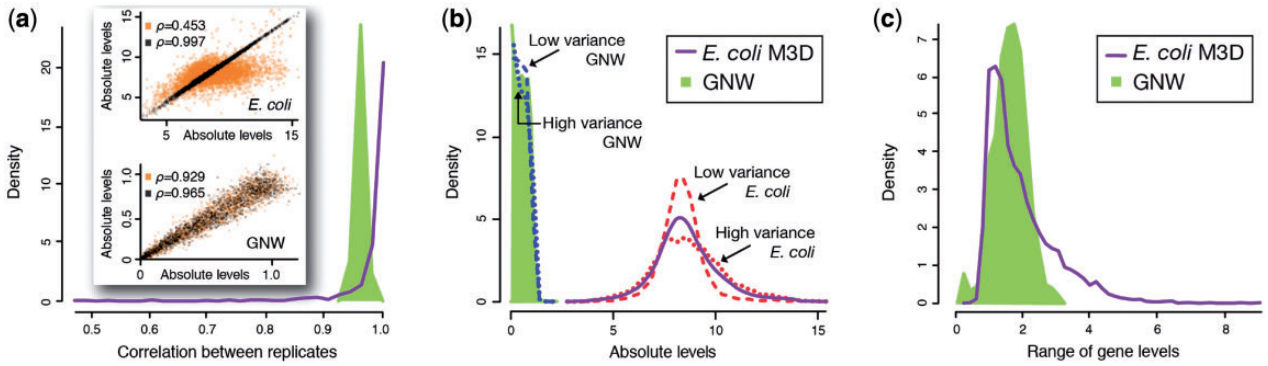

Fig. 2. Analysis of simple data histograms. Differences between real (*E.coli*: outline) and *in silico* (GNW: filled) data can be highlighted by the comparison of histograms. (**a**) Histogram of correlation coefficients computed between pairs of replicate microarrays (abscissa). Each inset panel scatter-plots absolute expression levels of the two pairs of replicates exhibiting the best (dark dots) and worst (light dots) correlation for *E.coli* (top) as well as GNW (bottom). (**b**) Overall histogram of intensities in a compendium. From the entire set of microarrays sorted by the variance of their intensities, a low (5% quantile of variance) and a high variance (95% quantile) microarray were selected that are shown as dashed and dotted lines, respectively. In contrast, (**c**) depicts intensity ranges for individual genes, i.e. differences between the observed minimum (5% quantile) and maximum (95% quantile) expression for each gene

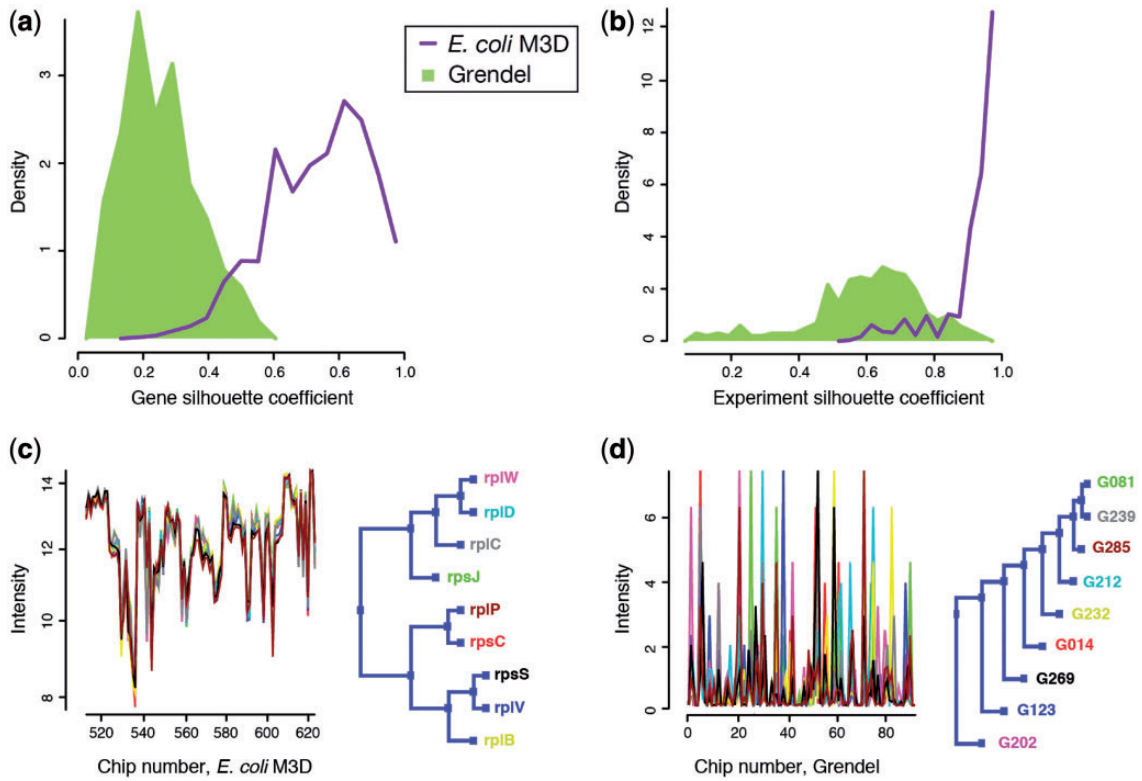

Fig. 3. Analysis of cluster quality. We grouped (**a**) genes and (**b**) microarrays by hierarchical clustering to examine the cluster quality via silhouette coefficients across all possible clusters. Cluster quality is consistently higher in real (here: *E.coli*, outline) than in artificial data (here: GRENDEL, filled). Individual gene clusters and corresponding gene profiles (see dendrogram for gene symbols matching the line color) are shown for two clusters that exhibit the highest silhouette coefficients, i.e. (**c**) for genes that code for ribosomal proteins in *E.coli* and (**d**) for artificial genes. Gene clusters exhibit substantially tighter expression patterns in case of real data, which is also reflected by topological differences between cluster dendrograms

from patterns of different biological functions (Eisen *et al.*, 1998). However, in artificial data, cluster quality measured by Silhouette coefficients is substantially lower both for gene and microarray clusters (Fig. 3a and b).

Co-expression patterns only emerge in artificial data for gene sets that are predominantly controlled by the same transcription factor. The fact that transcriptional regulation is the only mechanism to generate co-regulation patterns might explain that they are less pronounced in artificial data and form unusual cluster structures (Fig. 3c and d).

Similar to section 4.1, SynTReN yields a somewhat more realistic histogram for microarrays, while the gene cluster histogram is multi-modal and unlike the one shown for the real data. Again, the deviations between real data and GNW or GRENDEL are similar in both degree and shape.

4.3 Network analyses (SynTReN)

TF and TG dependencies. This and the subsequent section evaluate properties of GRIs. GRIs can be extracted from expression data in a simple but effective manner based on the correlations between TFs and TGs or between TGs and TGs (Butte and Kohane, 1999). In turn, the feasibility of such an approach can be assessed by examining whether pairs with high correlation values are actually enriched in true TF–TG or TG–TG dependencies (Fig. 4a and b). The comparison between the histograms derived from artificial and real data revealed two important differences. First, the enrichment of GRIs at high correlation values was markedly stronger for artificial data indicating that GRI inference is easier than in case of *E.coli* data. Second, artificial data showed similar degrees of enrichment for positive and negative correlations (activating and inhibiting relationships), whereas we could only detect an enrichment for positive correlations in case of real data (Fig. 4a). Apparently, mostly activating relationships can be observed in the real data, which is consistent with earlier observations (Naeem *et al.*, 2012).

TF activity histogram. The activity of transcriptional regulators such as TFs can be determined via statistical tests from expression changes exhibited by their targets (Fig. 4c). Strikingly, while the majority of TFs were found active in at least 10% of the real measurements, 50% of the TFs were never active in artificial profiles. Changes in TF activity are caused by transcriptional regulation only as other layers of regulation that could mediate such activity changes (e.g. protein modification) are neglected in current simulators.

Note that such gene set enrichment tests are used analogously for the detection of active TFs as well as of differentially expressed biological processes (Naeem *et al.*, 2012). As biological processes are typically not modeled in simulated datasets, we can still examine the performance of gene set enrichment tests via TF activities.

For the network analyses described in this section, the remaining two simulators exhibit similar patterns of deviation that are even stronger as those shown in Figure 4.

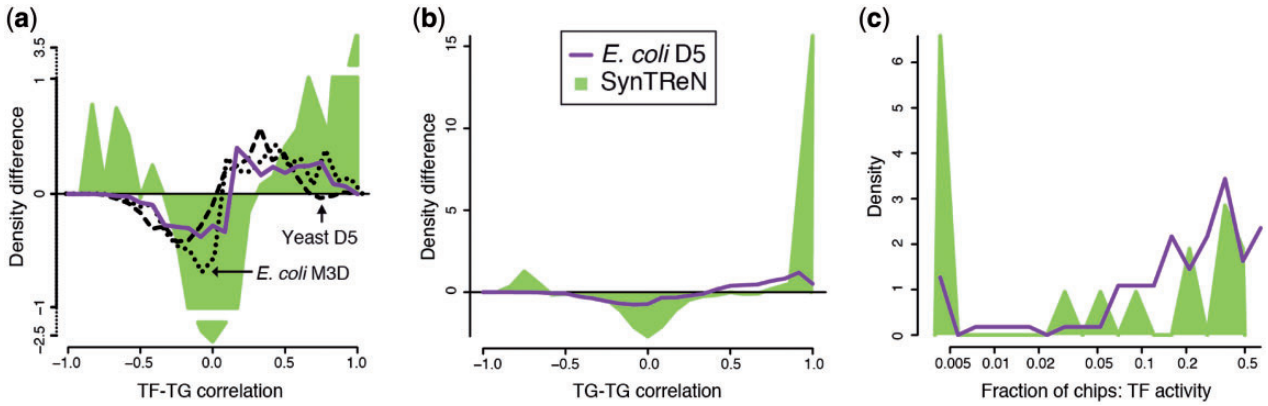

Fig. 4. Analysis of gene regulatory networks. Histograms are also applied to evaluate properties of TF–TG interactions in regulatory networks of *E.coli* and SynTReN. In (**a**) and (**b**), we examine to what extent TF–TG interactions and pairs of TGs regulated by the same TFs, respectively, can be identified based on correlation. Positive or negative differences in density [ordinate, non-continuous in (a)] indicate that the chances to detect true TF–TG or TG–TG relationships in the given correlation intervals are increased or decreased, respectively. Interactions were also analyzed for a yeast and an alternative *E.coli* dataset. In (**c**), the activity of TFs is inferred from the levels of differential expression exhibited by their TGs. Shown is the density (ordinate) of TFs active in a given fraction of the available chips. The leftmost histogram bin corresponds to a fraction of 0 and thus denotes TFs that are never found active

4.4 Organism comparison

Furthermore, we repeated the above analyses for a comparison between four real expression compendia compiled for *E.coli* and *S.cerevisiae* (Fig. 1, heatmap). Generally, the compared histograms exhibited differences that are substantially smaller than the deviations exhibited in the comparisons of real and artificial data described above. Although it is known that network inference is more difficult in eukaryotes (Küffner *et al.*, 2012; Marbach *et al.*, 2012; Narendra *et al.*, 2011), there are only little differences in the overall shape of *E.coli* and yeast histograms for TF–TG dependencies. However, in contrast to *E.coli*, a high correlation between TFs and TGs alone is not a good indicator for GRIs in yeast (Fig. 4a). This can also be observed in the heatmap of Figure 1, where the three examined network properties show a similar degree of deviation between *E.coli* and *S.cerevisiae* on the one hand and the real and artificial datasets on the other hand.

4.5 Separability of real and artificial datasets

We tested several scoring schemes to show that differences between the property histograms of real and artificial datasets are statistically significant based on the *t*-test. Real and artificial data are significantly different for 10 of 12 property histograms. In contrast, alternative scoring schemes, such as Wilcoxon's test or the relative entropy, could distinguish datasets for substantially fewer properties. See section 1 of the Supplementary Material for more details.

5 DISCUSSION

Simulated datasets aim to bridge the gap to dependable gold standards that are required for the development and validation of tools across many expression data analysis steps. For the application to algorithm benchmarks, simulators should be able to emulate the properties of real data. To test this ability, we applied several important steps in gene expression data analysis ranging from clustering to the inference of GRIs. The results of these analyses were represented as property histograms for the comparison of real and artificial datasets. Data generated by three gene network simulators were compared with four expression compendia measured in *E.coli* and *S.cerevisiae*.

We showed that property histograms facilitate the characterization and comparison of expression compendia. Most importantly, histograms were highly consistent across different real expression compendia and even between different organisms. In contrast, simulated data exhibited markedly different ranges of values and shapes. The captured properties thus clearly separated real and simulated data in analogy to Turing's test.

We observed strong deviations even in basic intensity and noise histograms. These are likely due to simplified simulation models omitting steps that play a role in real measurements such as the modeling of mRNA concentrations and their quantification by optical detectors. We constructed further histograms from clustering approaches to compare resulting clusters of genes and experimental conditions. Artificial co-regulation patterns as detected by clustering are exclusively generated by transcriptional regulation, as this is typically the only regulatory mechanism implemented in current simulators. However, the most striking and reproducible (i.e. found in all examined simulators) deviations were observed for properties derived from gene regulatory networks. This is surprising, as the examined simulators were particularly devised for the assessment of network inference approaches.

Many simulators aim to use realistic gene regulatory network topologies, e.g. sampled from experimentally derived GRIs, instead of random networks. However, other important layers of regulation such as different cellular states or signal transduction are neglected that in real data contribute to the complexity of expression patterns. Furthermore, executable models are constructed by arbitrarily selecting model parameters for the dynamic simulation of GRIs. Datasets generated in such a setup will differ substantially from real measurements even if realistic GRI topologies are used. Indeed, none of the different topologies

implemented by the three simulators substantially reduced the observed differences to real data.

Realistic data simulation requires a careful analysis of properties of network topologies, dynamic parameters and their relationship to the generated expression data. Our results suggest several improvements: (i) The expression model should be improved to produce normally distributed data in the range of typical expression measurement platforms, so that, for instance, fold changes have a similar meaning as in real datasets. The simulation of noise is important, but excessive noise levels should be avoided. (ii) The generation of realistic co-regulation patterns requires the emulation of groups of TFs, their combinatorial effects and the conditions for becoming active. (iii) As a related topic, we note that random parameterization (GNW, GRENDEL) or random mapping of experimentally determined parameters onto the network (SynTReN) is not adequate. Instead, an iterative optimization of model parameters could preserve dataset properties such as those that we described here. (iv) Simulators should implement regulation on the protein level. Currently, the simulation of just transcriptional regulation over-simplifies network inference, as all regulatory events become visible at the level of gene expression. (v) Finally, simulators should account for the diversity of genes, for instance, regarding the number of regulatory states, instead of assuming that one model is sufficient to generate all data. This would help to reproduce the broader distributions of properties derived from real data.

Our approach (see availability for Web site and software) offers a detailed comparative analysis of expression data generated by real and artificial gene regulatory networks. Our assessment is based on typical data processing and analysis pipelines. We thereby provided an important layer of assessment on top of data simulators that are essential tools for both data analysis and network inference when no reliable gold standards are available. The analysis suggested several ways for improving current simulators to generate more realistic datasets.

ACKNOWLEDGEMENTS

The authors thank Tobias Petri for discussion and advice.

Funding: This work has been supported by BMBF (http://www.bmbf.de, FKZ 01GS0801). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Albers,C.J. *et al.* (2006) SIMAGE: simulation of DNA-microarray gene expression data. *BMC Bioinformatics*, **7**, 205.

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

Butte,A.J. and Kohane,I.S. (1999) Unsupervised knowledge discovery in medical databases using relevance networks. *Proc. AMIA Symp.*, 711–715.

Draghici,S. *et al.* (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Faith,J.J. *et al.* (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.

Fundel,K. *et al.* (2008) Normalization and gene p-value estimation: issues in microarray data processing. *Bioinform. Biol. Insights*, **2**, 291–305.

Gama-Castro,S. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.

Haynes,B.C. and Brent,M.R. (2009) Benchmarking regulatory network reconstruction with GRENDEL. *Bioinformatics*, **25**, 801–807.

Hebenstreit,D. and Teichmann,S.A. (2011) Analysis and simulation of gene expression profiles in pure and mixed cell populations. *Phys. Biol.*, **8**, 035013.

Küffner,R. *et al.* (2010) Petri Nets with Fuzzy Logic (PNFL): reverse engineering and parametrization. *PLoS One*, **5**, e12807.

Küffner,R. *et al.* (2012) Inferring gene regulatory networks by ANOVA. *Bioinformatics*, **28**, 1376–1382.

MacIsaac,K.D. *et al.* (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, **7**, 113.

Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Naeem,H. *et al.* (2012) Rigorous assessment of gene set enrichment tests. *Bioinformatics*, **28**, 1480–1486.

Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief Bioinform.*, **9**, 189–197.

Narendra,V. *et al.* (2011) A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. *Genomics*, **97**, 7–18.

Needham,C.J. *et al.* (2006) Inference in Bayesian networks. *Nat. Biotechnol.*, **24**, 51–53.

Parrish,R.S. *et al.* (2009) Distribution modeling and simulation of gene expression data. *Comput. Stat. Data Anal.*, **53**, 1650–1660.

Pinna,A. *et al.* (2011) Simulating systems genetics data with SysGenSIM. *Bioinformatics*, **27**, 2459–2462.

Rocke,D.M. *et al.* (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics*, **25**, 701–702.

Rousseeuw,P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Schaffter,T. *et al.* (2011) GeneNetWeaver: *in silico* benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**, 2263–2270.

Van den Bulcke,T. *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, **7**, 43.

Wu,M. and Chan,C. (2012) Learning transcriptional regulation on a genome scale: a theoretical analysis based on gene expression data. *Brief Bioinform.*, **13**, 150–161.

Wu,S. *et al.* (2004) Cluster analysis of gene expression data based on self-splitting and merging competitive learning. *IEEE Trans. Inf. Technol. Biomed.*, **8**, 5–15.

Yeung,K.Y. *et al.* (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.