Journal of Bioinformatics and Computational Biology
Vol. 6, No. 3 (2008) 493–519

Imperial College Press
www.icpress.co.uk

USING DIRECTED INFORMATION TO BUILD BIOLOGICALLY RELEVANT INFLUENCE NETWORKS

ARVIND RAO* and ALFRED O. HERO III†

Department of Electrical Engineering and Computer Science and Department of Bioinformatics, University of Michigan Ann Arbor, MI 48109, USA
**ukarvind@umich.edu*
†hero@umich.edu

DAVID J. STATES

Department of Bioinformatics and Department of Human Genetics University of Michigan, Ann Arbor, MI 48109, USA
dstates@umich.edu

JAMES DOUGLAS ENGEL

Department of Cell and Developmental Biology University of Michigan, Ann Arbor, MI 48109, USA
engel@umich.edu

Received 1 August 2007
Revised 1 December 2007
Accepted 3 January 2008

The systematic inference of biologically relevant influence networks remains a challenging problem in computational biology. Even though the availability of high-throughput data has enabled the use of probabilistic models to infer the plausible structure of such networks, their true interpretation of the biology of the process is questionable. In this work, we propose a network inference methodology, based on the directed information (DTI) criterion, that incorporates the biology of transcription within the framework so as to enable experimentally verifiable inference. We use publicly available embryonic kidney and T-cell microarray datasets to demonstrate our results. We present two variants of network inference via DTI — supervised and unsupervised — and the inferred networks relevant to mammalian nephrogenesis and T-cell activation. Conformity of the obtained interactions with the literature as well as comparison with the coefficient of determination (CoD) method are demonstrated. Apart from network inference, the proposed framework enables the exploration of specific interactions, not just those revealed by data. To illustrate the latter point, a DTI-based framework to resolve interactions

*Corresponding author.

between transcription factor modules and target coregulated genes is proposed. Additionally, we show that DTI can be used in conjunction with mutual information to infer higher-order influence networks involving cooperative gene interactions.

Keywords: Mutual information; directed information; transcription factor module; comparative genomics; transcription regulatory network.

1. Introduction

Computational methods for inferring dependencies between genes[1–3] using probabilistic techniques have been used for quite some time now. However, the biological significance of these recovered networks has been a topic of debate, apart from the fact that such approaches mostly yield networks of significant influences as observed/inferred from the underlying structure of data. Alternatively, other biological data (such as sequence information) might suggest the examination of the probabilistic dependence of one gene on another gene through the transcription factor (TF) encoded by the first gene. What if we were interested in the transcriptional influences on a certain gene A, but our prospective network inference technique was unable to recover them? We propose a technique with an eye on two of these challenges: biological significance and influence determination between any two variables of interest. Such an approach is increasingly necessary in order to integrate and understand multiple sources of data (sequence, expression, etc.).

The method that we propose builds on an information theoretic criterion referred to as the directed information (DTI). DTI is a variant of mutual information (MI) that attempts to capture the direction of information flow. It is widely used in the analysis of communication systems with feedback or feedforward[4–6] as well as in economic time series analysis.[6,7] DTI[4,8] can be interpreted as a directed version of mutual information, a criterion used quite frequently in other related works.[1] It turns out, as we will demonstrate, that DTI gives a sense of directional association for the principled discovery of biological influence networks.

The contributions of this work are as follows. Firstly, we present a short theoretical treatment of DTI and an approach to the supervised and unsupervised discovery of influence networks, using microarray expression data. Secondly, we examine two scenarios — the inference of large-scale gene influence networks (in mammalian nephrogenesis and T-cell development) as well as potential effector genes for *Gata3* transcriptional regulation in distinct biological contexts. We find that this method outperforms other methods in several aspects and leads to the formulation of biologically relevant hypotheses that might aid subsequent experimental investigation. Finally, we conclude with the application of DTI to two important questions in bioinformatics: TF module discovery and higher-order network inference. TF module discovery is the identification of common regulatory modules (groups of TFs) whose binding sites co-occur on the promoters of coexpressed genes. Higher-order network inference, in this work, examines the resolution of three-way interactions rather than only pairwise relationships among genes.[9]

2. Organization

This paper is organized as follows. In Sec. 3, the working definition of transcriptional gene networks is given. Based on this definition, four main research problems are posed — those pertaining to supervised and unsupervised network inference, TF module–gene interactions, and inference of higher-order influence networks. DTI is proposed as part of a general framework to answer these questions (Sec. 5), and a methodology for the determination of influence and its significance is examined (Appendix and Sec. 6). The paper concludes with results applicable to each of the questions posed above (Sec. 8), using a combination of synthetic and real biological data.

3. Gene Networks

Transcription is the process of generating messenger RNA (mRNA) from the DNA template representing the gene. It is the intermediate step before the generation of functional proteins from messenger RNA. During gene expression (Fig. 1), TF proteins are recruited at the proximal promoter of the gene as well as at distal sequence elements (enhancers/silencers) which can lie several hundreds of kilobases from the gene's transcriptional start site.[10] Since TFs are also proteins (or their activated forms), which in turn are encoded for by other genes, the notion of an influence network between a TF gene and the target gene can be considered.

In Fig. 2, a characterization of transcriptional regulatory networks, as relevant to this work, is given. As the name suggests, gene A is connected by a link to gene C if a product of gene A, say protein A, is involved in the transcriptional regulation of gene C. This might mean that protein A is involved in the formation of the complex, which binds at the basal transcriptional machinery of gene C to drive gene C regulation.

As can be seen, the components of the TF complex recruited at the gene promoter are the products of several genes. Therefore, the incorrect inference of a transcriptional regulatory network can lead to false hypotheses about the actual

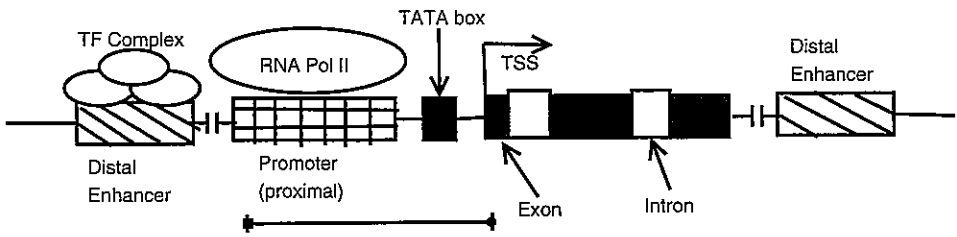

Fig. 1. Schematic of transcriptional regulation. Sequence motifs at the promoter and the distal regulatory elements together confer specificity of gene expression via TF binding.

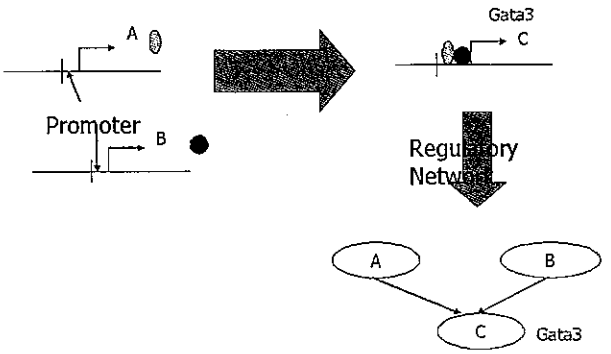

Fig. 2. A transcriptional regulatory network with genes A and B effecting C. An example of C that we study here is the *Gata3* gene.

set of genes affecting a target gene. Since biologists are increasingly relying on computational tools to guide experiment design, a principled approach to biologically relevant network inference can lead to significant savings in time and resources in downstream experimental design. In this paper, we try to combine some of the other available biological data (phylogenetic conservation of binding sites across genomes and expression data) to build network topologies with a lower false-positive rate of linkage. Some previous work in this regard has been reported in Mac Isaac and Fraenkel[11] and in Kreiman.[12]

4. Problem Setup

In this work, we also study the mechanism of gene regulation, with the *Gata3* gene as an example. This gene has important roles in several processes in mammalian development,[10,13] such as in the developing urogenital system (nephrogenesis), central nervous system, and T-cell development. In order to find out which TFs regulate the tissue-specific transcription of *Gata3* (either at the promoter or long-range regulatory elements), a commonly followed approach[11,12] is to look for phylogenetically conserved transcription factor binding sites (TFBSs). The hypothesis underlying this strategy is that the interspecies conservation of a TFBS suggests a possibly functional binding of the TF at the motif (from evolutionary pressure for function). With a view to understanding gene regulatory mechanisms, this work primarily addresses the following issues:

- Biologists are also interested in the network of relationships among genes expressed under a certain set of conditions that uses several network inference procedures such as Bayesian networks,[3] mutual information,[1] etc. However, there has been a lack of a common framework to do both supervised and unsupervised directed network inference within these settings in order to detect directed nonlinear gene–gene interactions. We present DTI as a potential solution in both these scenarios. Supervised network inference pertains to finding the strengths of

directed relationships between two specific genes. Unsupervised network inference deals with finding the most probable network structure to explain the observed data (like in Bayesian structure learning using expression data). This is addressed in Secs. 8.2 and 8.3.

- Which TFs are potentially active at the target gene's promoter during its tissue-specific regulation? This question is primarily answered by examining the phylogenetically conserved TFBSs at the promoter, and asking if microarray expression data suggest the presence of an influence between the TF-encoding gene and the target gene (i.e. *Gata3*). This approach thus integrates sequence and expression information (Sec. 8.4).
- Which TFs underlie the tissue-specific expression of a group of coexpressed/coregulated genes (e.g. *Gata3* and others)? One common approach is to search the proximal promoters of all such tissue-specific genes and to look for modules of TFs that control tissue-specific expression.[11,12] For the *Gata3* example, we ask if there are any TFs underlying ureteric bud (UB)-specific expression for *Gata3* during nephrogenesis. For this purpose, we find modules from coexpressed gene promoters and use microarray expression to point out possible effectors of target gene expression (Sec. 8.5).
- Gene interactions during processes such as development and disease progression are rarely pairwise, and occur in cliques such as pathways. Additionally, cross-talk between components of different pathways is essential in the progression of such dynamic processes. To this end, the inference of higher-order interactions (more than only two-way gene relationships) is seen to be a useful approach.[9] Using DTI, it would be interesting to find directed interactions between differentially expressed genes of the developing kidney to determine pathway cross-talk (Sec. 8.6).

4.1. *Phylogenetic conservation of transcription factor binding sites (TFBSs)*

As mentioned above, the mechanism of regulation of a target gene is via the binding site of the corresponding TF. It is believed that several TF binding motifs might have appeared over the evolutionary time period due to insertions, mutations, deletions, etc. in vertebrate genomes. However, if we are interested in the regulation of a process which is known to be similar between several organisms (say human, chimpanzee, mouse, rat, and chicken), then we can look for the conservation of functional binding sites over all of these genomes. This helps us isolate the putatively functional binding sites, as opposed to those which might have randomly arisen; this, however, does not suggest that those other TF binding sites have no functional role. If we are interested in the mechanism of regulation of the *Gata3* gene (which is known to be implicated in mammalian nephrogenesis), we examine its promoter region for phylogenetically conserved TFBSs (Fig. 3). Such information can be obtained from most genome browsers.[14] We see that even for a fairly short

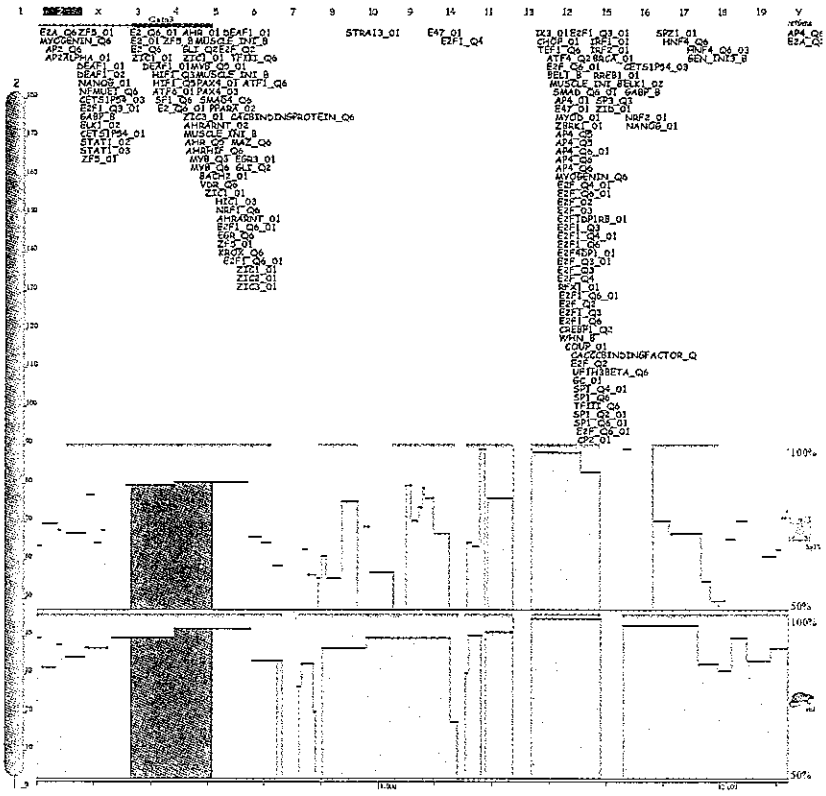

Fig. 3. TFBS conservation between human, mouse, and rat, upstream (x-axis) of *Gata3*. Source: http://www.ecrbrowser.dcode.org/.

stretch of sequence (1 kb) upstream of the gene, there are several conserved sequence elements which are potential TFBSs (light-gray regions in Fig. 3).

In Fig. 3, we have aligned the mouse *Gata3* promoter region with its human and rat counterparts. The height of each of the dark-gray regions indicates the extent of conservation between these species. Furthermore, it indicates that several TFs bind at these conserved regions. To test their functional role *in vivo* or *in vitro*, it is necessary to select only a subset of these TFs because of the great reliance on resources and effort. Hence, the genes coding for these conserved TFs are the ones that we examine for possible influence determination via expression-based influence metrics. If we are able to infer an influence between the TF-coding gene and the target gene at which its TF binds, then this reduces the number of candidates to be tested. To examine *Gata3*'s role in kidney development, we use microarray expression data from a public repository of kidney microarray data (http://genet.chmcc.org/, http://spring.imb.uq.edu.au/, and http://kidney.scgap.org/index.html). Each of these sources contains expression data profiling kidney development from about day 10.5 dpc to the neonate stage. Some of these studies also examine expression in

the developing ureteric bud (UB) and metanephric mesenchyme (MM) apart from the whole kidney.

Our approach thus integrates several aspects:

- using phylogenetic information to infer which TFBSs upstream of a target gene may be functional; and
- identifying if any of the TF genes influence a target gene by coding for a TF that binds at the site discovered from conservation studies. This directed influence is captured using an influence metric (e.g. DTI) in conjunction with expression data,[15,16] and is explained in Sec. 5.

5. DTI Formulation

As alluded to above, there is a need for a viable influence metric that can find relationships between the TF effector gene (identified from phylogenetic conservation) and the target gene (e.g. *Gata3*). Several such metrics have been proposed, notably, correlation, coefficient of determination (CoD), mutual information, etc. To alleviate the challenge of detecting nonlinear gene interactions, an information theoretic measure like mutual information has been used to infer the conditional dependence among genes by exploring the structure of the joint distribution of the gene expression profiles.[1] However, the absence of a directed dependence metric has hindered the utilization of the full potential of information theory. In this work, we examine the applicability of one such metric — the directed information (DTI) criterion — for the inference of nonlinear, directed gene influences.

DTI, which is a measure of the directed dependence between two N-length random processes $X \equiv X^N$ and $Y \equiv Y^N$, is given by[5]

$$I(X^N \to Y^N) = \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1}), \tag{1}$$

where Y^n denotes $(Y_1, Y_2, \ldots, Y_n)$, i.e. a segment of the realization of a random process Y, and $I(X^N; Y^N)$ is the Shannon mutual information.[17]

An interpretation of the above formulation for DTI is in order. To infer the notion of influence between two time series (mRNA expression data), we find the mutual information between the entire evolution of gene X (up to the current instant n) and the current instant of Y (Y_n), given the evolution of gene Y up to the previous instant $n-1$ (i.e. Y^{n-1}). This is done for every instant, $n \in (1, 2, \ldots, N)$, in the N-length expression time series.

As already known, $I(X^N; Y^N) = H(X^N) - H(X^N|Y^N)$, with $H(X^N)$ and $H(X^N|Y^N)$ being the entropy of X^N and the conditional entropy of X^N given Y^N, respectively. Using this definition of mutual information, DTI can be expressed in terms of individual and joint entropies of X^N and Y^N. The task of N-dimensional entropy estimation is an important one, but due to computational complexity and moderate sample size, histogram estimation of multivariate density is unviable.

Nevertheless, several methods exist for consistent entropy estimation of multivariate small sample data.[18–21] In the context of microarray expression data, wherein probe-level and technical/biological replicates are available, we use the method of Miller[18] for entropy estimation.

From Eq. (1), we have

$$I(X^N \to Y^N) = \sum_{n=1}^{N} [H(X^n|Y^{n-1}) - H(X^n|Y^n)]$$
$$= \sum_{n=1}^{N} \{[H(X^n, Y^{n-1}) - H(Y^{n-1})] - [H(X^n, Y^n) - H(Y^n)]\}. \quad (2)$$

- To evaluate the DTI expression in Eq. (2), we need to estimate the entropy terms $H(X^n, Y^{n-1})$, $H(Y^{n-1})$, $H(X^n, Y^n)$, and $H(Y^n)$. This involves the estimation of marginal and joint entropies of n random variables, each of which is R-dimensional, R being the number of replicates (probe-level, biological, and technical).
- Although some approaches need the estimation of probability density of the R-dimensional multivariate data (X^n) prior to entropy estimation, one way to circumvent this is to use the method proposed in Miller.[18] This approach uses a Voronoi tessellation of the R-dimensional space to build nearly uniform partitions (of equal mass) of the density. The set of Voronoi regions $(V^1, V^2, \ldots, V^n)$ for each of the n points in R-dimensional space is formed by associating, with each point X_k, a set of points V^k that are closer to X_k than any other point X_l, where the subscripts k and l pertain to the kth and lth time instants of gene expression, respectively.
- Thus, the entropy estimator is expressed as $\hat{H}(X^n) = \frac{1}{n}\sum_{i=1}^{n} \log(nA(V^i))$, where $A(V^i)$ is the R-dimensional volume of the Voronoi region V^i. $A(V^i)$ is computed as the area of the polygon formed by the vertices of the convex hull of the Voronoi region V^i. This estimate has low variance and is asymptotically efficient.[22]

To obtain the DTI between any two genes of interest (X and Y) with N-length expression profiles X^N and Y^N, respectively, we plug in the entropy estimates computed above into Eq. (2).

From the definition of DTI, we know that $0 \leq I(X_i^N \to Y^N) \leq I(X_i^N; Y^N) < \infty$. For easy comparison with other metrics, we use a normalized DTI metric (see Appendix) given by $\rho_{\mathrm{DTI}} = \sqrt{1 - e^{-2I(X^N \to Y^N)}} = \sqrt{1 - e^{-2\sum_{i=1}^{N} I(X^i; Y_i | Y^{i-1})}}$. This maps the large range of DTI, $([0, \infty])$, to lie in $[0, 1]$. Another point of consideration is to estimate the significance of the true DTI value compared to a null distribution on the DTI value (i.e. what is the chance of finding the DTI value by coincidence from the series X and Y?). This is done using empirical p-value estimation after bootstrap resampling (Sec. 6). A threshold p-value of 0.05 is used to estimate the significance of the true DTI value in conjunction with the density of a random data permutation, as outlined below.

6. Significance Estimation of DTI

We now outline a procedure to estimate the empirical p-value in order to ascertain the significance of the normalized directed information $\hat{I}(X^N \to Y^N)$ between any two N-length time series $X \equiv X^N = (X_1, X_2, \ldots, X_N)$ and $Y \equiv Y^N = (Y_1, Y_2, \ldots, Y_N)$. In our case, the detection statistic is $\Theta = \hat{I}(X^N \to Y^N)$ and the chosen acceptable p-value is α.

The overall bootstrap-based test procedure is as follows[23–25]:

- Repeat the following procedure $B(= 1{,}000)$ times (with index $b = 1, 2, \ldots, B$):
  - — Generate resampled (with replacement) versions of the time series X^N, Y^N, denoted by X_b^N and Y_b^N, respectively.
  - — Compute the statistic $\theta^b = \hat{I}(X_b^N \to Y_b^N)$.
- Construct an empirical cumulative distribution function (CDF) from these bootstrapped sample statistics, as $F_\Theta(\theta) = P(\Theta \leq \theta) = \frac{1}{B}\sum_{b=1}^{B} I_{x \geq 0}(x = \theta - \theta^b)$, where I is an indicator random variable on its argument x.
- Compute the true detection statistic (on the original time series) $\theta_0 = \hat{I}(X^N \to Y^N)$ and its corresponding p-value ($p_0 = 1 - F_\Theta(\theta_0)$) under the empirical null distribution $F_\Theta(\theta)$.
- If $F_\Theta(\theta_0) \geq (1 - \alpha)$, then we have that the true DTI value is significant at level α, leading to rejection of the null hypothesis (no directional association).

7. Summary of Algorithm

We now present two versions of the DTI algorithm: one which involves an inference of general influence network between all genes of interest (unsupervised DTI); and another, a focused search for effector genes which influence one particular gene of interest (supervised DTI).

Our proposed approach using (supervised DTI) for determining the effectors for gene B is as follows:

- Identify the G genes $(A_1, A_2, \ldots, A_G)$, based on required phenotypical characteristics, using fold change studies. Preprocess the gene expression profiles by normalization and cubic spline interpolation. Assuming that there are N points for each gene, entropy estimation is used to compute the terms in the DTI expression [Eq. (2)].
- For each pair of genes A_i and B among these G genes,
  - — Look for a phylogenetically conserved TFBS encoded by gene A_i in the upstream region of gene B.
  - — Find $DTI(A_i, B) = I(A_i^N \to B^N)$, and the normalized DTI from A_i to B, $\rho_{\mathrm{DTI}}(A_i, B) = \sqrt{1 - e^{-2I(A_i^N \to B^N)}}$.
  - — Bootstrap resampling over the data points of A_i and B yields a null distribution for $DTI(A_i, B)$. If the true $DTI(A_i, B)$ is greater than the 95% upper

limit of the confidence interval (CI) from this null histogram, infer a potential influence from A_i to B.
— The value of the normalized DTI from A_i to B gives the putative strength of interaction/influence.
— Every gene A_i which is potentially influencing B is an effector. This search is done for each gene A_i among these G genes $(A_1, A_2, \ldots, A_G)$.

As can be seen, phylogenetic information is inherently built into the influence network inference step above. We note that, in supervised DTI, the choice of potential effectors for a target gene is based on only those TFs that have a binding site at the target gene's promoter. In this sense, supervised DTI aims to reduce the overall search space based on biological prior knowledge.

For unsupervised DTI, we adapt the above approach for every pair of genes (A, B) in the list, noting that $DTI(A, B) \neq DTI(B, A)$. In this case, we are not looking at any interaction in particular, but are interested in the entire influence network that can be potentially inferred from the given time series expression data. The network adjacency matrix has entries depending on the direction of influence, and is related to the strength of influence as well as control of false discovery rate (FDR). The Benjamini–Hochberg procedure[26] is used to screen each of the $M(= G(G-1))$ hypotheses (both directions) during network discovery among G genes.

Briefly, the FDR procedure controls the expected proportion of false positives among the total number of rejections rather than just the chance of false positives.[27] It tolerates more false positives and allows fewer false negatives.

- The p-values of the various edges $(1, 2, \ldots, M)$ are ranked from lowest to highest, all satisfying the original significance cut-off of $p = 0.05$. The ranked p-values are designated as $p_{(1)}, p_{(2)}, \ldots, p_{(M)}$.
- For $j = 1, 2, \ldots, M$, the null hypothesis (no edge) H_j is rejected at level α if $p_{(j)} \leq \frac{j}{M}\alpha$.
- All of the edges with p-value $\leq p_{(j)}$ are retained in the final network.

In Table 1, we compare the various contemporary methods of directed network inference. Recent literature has introduced several interesting approaches such as graphical Gaussian models (GGMs), coefficient of determination (CoD), and state space models (SSMs) for directed network inference. This comparison is based primarily on expectations from such inference procedures — that we would like any such metric/procedure to

- resolve cycles in recovered interactions;
- be capable of resolving directional and potentially nonlinear interactions. This is because interactions among genes involve nonlinear kinetics;
- be a nonparametric procedure to avoid distributional assumptions (e.g. noise); and

Table 1. Comparison of various network inference methods.

Method	Resolve cycles	Nonlinear framework	Search for interaction	Nonparametric framework
SSM[28,29]	Y	Y	N	Y
CoD[30]	N	N	Y	N
GGM[2]	N	Y	N	N
DTI[8]	Y	Y	Y	Y

Y, yes; N, no.

- be capable of recovering interactions that a biologist might be interested in. Instead of using a method that discovers interactions underlying the data purely, the biologist should be able to use prior knowledge (from the literature, perhaps). For example, a biologist can examine the strength and significance of a known interaction, and use this as a basis for finding other such interactions.

From the above comparisons, we see that DTI is one metric which can recover interactions under all of these considerations.

8. Results

In this section, we give some scenarios where DTI can complement existing bioinformatics strategies to answer several questions pertaining to transcriptional regulatory mechanisms. We address four different questions:

- To infer gene influence networks between genes that have a role in early kidney development and T-cell activation, we use unsupervised DTI with relevant microarray expression data, noting that these influence networks are not necessarily transcriptional regulatory networks.
- To find TFs that might be involved in the regulation of a target gene (like *Gata3*) at the promoter, a common approach is to first look for phylogenetically conserved TFBS sequences across related species. These species are selected based on whether the particular biological process is conserved in them. To add additional credence to the role of these conserved TFBSs, microarray expression can be integrated via supervised DTI to check for evidence of an influence between the TF-encoding gene and the target gene.
- Thirdly, we examine the promoters of several genes that have a documented role in ureteric bud (UB) development. The idea is to look for common TF modules that govern the combined coexpression and coregulation of these genes.[11] Again, expression data and supervised DTI can be used to check for influences between the module components and the target gene(s).
- Finally, the problem of inferring higher-order dependencies between various genes using a combination of mutual and directed information is presented in the context of differentially expressed UB-specific genes of the developing kidney.

Before proceeding, we examine the performance of this approach on synthetic data.

8.1. *Synthetic network*

A synthetic network is constructed in the following fashion. We assume that there are two genes, g_1 and g_3 (both of which are modeled as uniform random variables), which drive the remaining genes of a nine gene network. The evolution equations are as below:

$$
\begin{aligned}
g_{2,t} &= \frac{1}{2}g_{1,t-1} + \frac{1}{3}g_{3,t-2} + g_{7,t-1} \\
g_{4,t} &= g_{2,t-1}^2 + g_{3,t-1}^{1/2} \\
g_{5,t} &= g_{2,t-2} + g_{4,t-1} \\
g_{6,t} &= g_{4,t-1} + g_{2,t-2}^{1/2} \\
g_{7,t} &= \frac{1}{2}g_{4,t-1}^{1/3} \\
g_{8,t} &= \frac{1}{2}g_{6,t-1}^{1/3} + \frac{1}{3}g_{7,t-1}^{1/2} \\
g_{9,t} &= \frac{2}{3}g_{4,t-1}^{2/3} + \frac{1}{4}g_{7,t-2}^{1/2}.
\end{aligned}
$$

For the purpose of comparison, we study the performance of the CoD approach for directed influence network determination. The CoD allows the determination of association between two genes via an R^2 goodness-of-fit statistic. The methods of Hashimoto *et al.*[30] and Li *et al.*[31] are implemented on the time series data. Such a study would be useful to determine the relative merits of each approach. We believe that no one procedure can work for every application, and the choice of an appropriate method would be governed by the biological question under investigation. Each of these methods uses some underlying assumptions; if these are consistent with the question that we ask, then that method has utility.

As can be seen in Fig. 4, though CoD can detect linear lag influences, the strongly nonlinear ones are missed. DTI detects these influences and reproduces the synthetic network exactly. Given the nonlinear nature of transcriptional kinetics, this is essential for reliable network inference. DTI is also able to resolve loops and cycles ($g_3, [g_2, g_4], g_5$ and g_2, g_4, g_7, g_2). Based on these observations, we examine the networks inferred using DTI in both the supervised and unsupervised settings.

8.2. *Directed network inference:Gata3 regulation in early kidney development*

Biologists have an interest in influence networks that might be active during organ development. Advances in laser capture microdissection coupled with those in microarray methodology have enabled the investigation of temporal profiles of genes

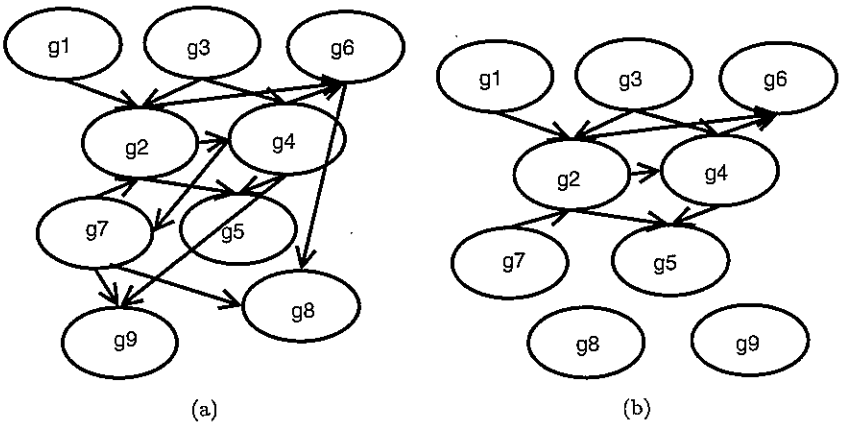

Fig. 4. The synthetic network as recovered by (a) DTI and (b) CoD.

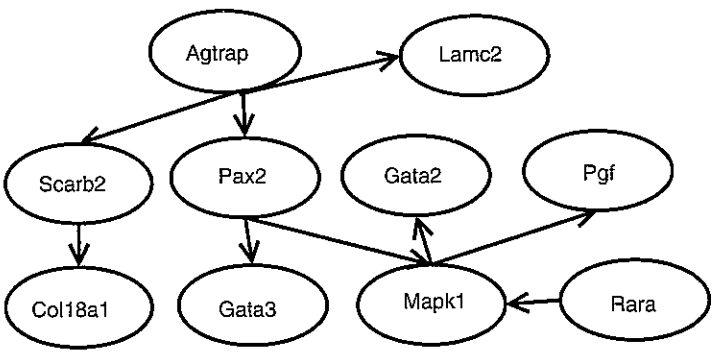

Fig. 5. Overall influence network using DTI during early kidney development.

putatively involved in these embryonic processes. Forty-seven genes are differentially expressed between the UB and MM,[16] and are putatively involved in bud branching during kidney development. The expression data[15] temporally profile kidney development from day 10.5 dpc to the neonate stage. The influence network among these genes is shown in Fig. 5. Several of the presented interactions are biologically validated, and there is an interest to confirm the novel ones pointed out in the network. The annotations of some of these genes are given in Table 2.

Some of the interactions that have been experimentally validated include the *Rara–Mapk1*,[32] *Pax2–Gata3*,[33] and *Agtr–Pax2*[34] interactions. We note that this result clarifies the application of DTI for network inference in an unsupervised manner, i.e. discovering interactions revealed by data rather than examining the strengths of interactions known *a priori*. Such a scenario will be explored later (Sec. 8.4). We note that, though several interaction networks are recovered, we only show the largest network including *Gata3* because this is the gene of interest in this study.

Table 2. Functional annotations (Entrez Gene) of some of the genes coexpressed with *Gata2* and *Gata3* during nephrogenesis.

Gene symbol	Gene name	Possible role in nephrogenesis (function)
Rara	Retinoic acid receptor	Crucial in early kidney development
Gata2	*GATA* binding protein 2	Several aspects of urogenital development
Gata3	*GATA* binding protein 3	Several aspects of urogenital development
Pax2	Paired homeobox-2	Conversion of MM precursor cells to tubular epithelium
Lamc2	Laminin	Cell adhesion molecule
Pgf	Placental growth factor	Arteriogenesis, growth factor activity during development
Col18a1	Collagen, type *XVIII*, alpha 1	Extracellular matrix structural constituent, cell adhesion
Agtrap	Angiotensin II receptor-associated protein	Ureteric bud cell branching

8.3. *Directed network inference: T-cell activation*

To clarify the validity of the presented approach, we present a similar analysis on another data set — the T-cell expression data[28] (Fig. 6). This data set represents the expression of various genes after T-cell activation using stimulation with phorbolester PMA and ionomycin. The data set contains the profiles of about 58 genes over 10 time points with 44 replicate measurements for each time point.

Several of these interactions have been confirmed in earlier studies,[28,35–37] and again point to the strength of DTI in recovering known interactions. The annotations of some of these genes are given in Table 3. We note that the network in Fig. 6 shows the largest influence network (containing *Gata3*) that can be recovered.

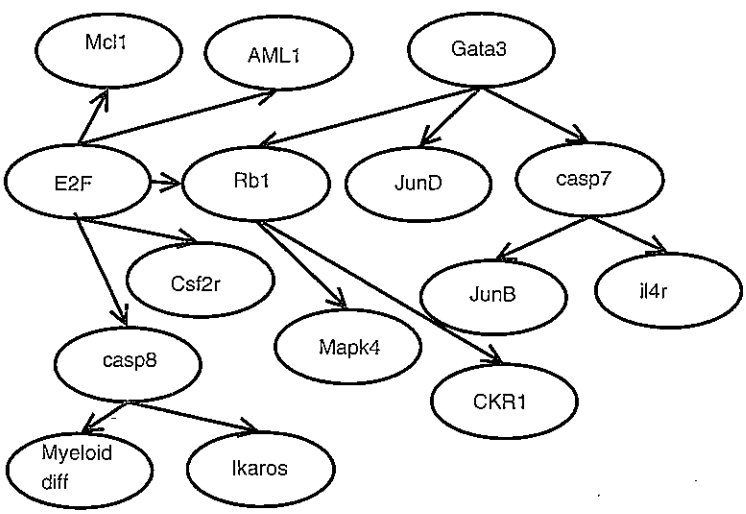

Fig. 6. DTI-based T-cell network.

Table 3. Functional annotations of some of the genes following T-cell activation.

Gene symbol	Gene name	Possible role in T-cell activation (function)
Casp7	Caspase 7	Involved in apoptosis
JunD	Jun D proto-oncogene	Regulatory role of in T lymphocyte proliferation and Th cell differentiation
CKR1	Chemokine receptor 1	Negative regulator of the antiviral $CD8^+$ T-cell response
Il4r	Interleukin 4 receptor	Inhibits *IL4*-mediated cell proliferation
Mapk4	Mitogen activated kinase 4	Signal transduction
AML1	Acute myeloid leukemia 1; aml1 oncogene	CD4 silencing during T-cell differentiation
Rb1	Retinoblastoma 1	Cell cycle control

Gata3 is involved in T-cell development as well as kidney development, and hence it is interesting to see networks relevant to each context in Figs. 5 and 6. Also, these 58 genes relevant to T-cell activation are very different from those for kidney development, with fairly low overlap. For example, this list does not include *Pax2* (which is relevant in the kidney development data).

8.4. *Phylogenetic conservation of TFBS effectors*

A common approach for the determination of functional TFBSs in genomic regions is to look for motifs in conserved regions across various species. Here, we focused on the interspecies conservation of TFBSs (Fig. 3) in the *Gata3* promoter to determine which of them might be related to transcriptional regulation of *Gata3*. Such a conservation across multiple species suggests selective evolutionary pressure on the region, with a potential relevance for function. As can be seen in Fig. 3, we examined the *Gata3* gene promoter and found at least 40 different TFs that could putatively bind at the promoter as part of the transcriptional complex. Some of these TFs, however, belong to the same family.

Using supervised DTI, we examined the strength of influence from each of the TF-encoding genes (A_i) to *Gata3*, based on expression level[15] (http://spring.imb.uq.edu.au/). These strength-of-influence DTI values were first checked for significance at a p-value of 0.05, and then ranked from highest to lowest (noting that the objective is to maximize $I(A_i \rightarrow Gata3)$). Based on this ranking, we indicate some of the TFs that have the highest influence on *Gata3* expression (Fig. 7). Obviously, this information is far from complete, because of examination only at the mRNA level for both effectors and *Gata3*.

Table 4 shows the embryonic kidney-specific expression of the TFs from Fig. 7. This is an independent annotation obtained from UniProt (http://expasy.org/sprot/). To understand the notion of kidney-specific regulation of *Gata3* expression by various TFs, we have integrated three different criteria: we expect that the TFs regulating expression would have an influence on *Gata3* expression, be expressed in the kidney, and have a conserved binding site at the *Gata3* promoter. This is clarified in part by Fig. 7 and Table 4. As an example, we see that the TFs *Pax2*,

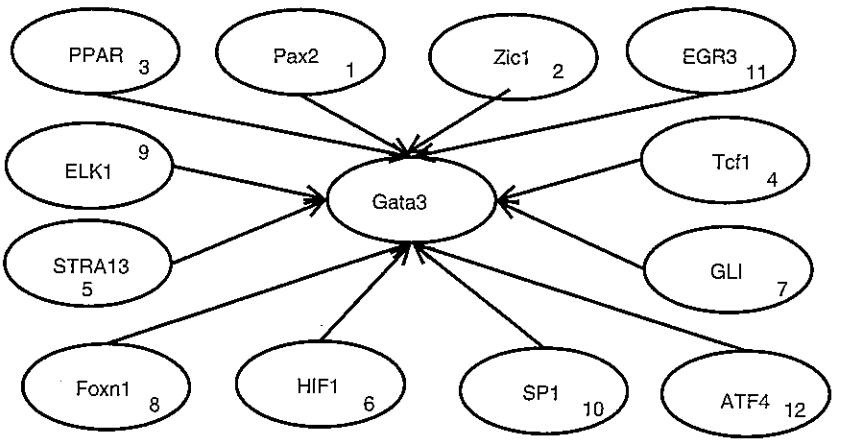

Fig. 7. Putative upstream TFs using DTI for the *Gata3* gene. The numbers in each TF oval represent the DTI rank of the respective TF.

Table 4. Functional annotations of some of the TF genes putatively influencing *Gata3* regulation in kidney.

Gene symbol	Description	Expressed in kidney
PPAR	Peroxisome proliferator-activated receptor	Yes
Pax2	Paired homeobox-2	Yes
HIF1	Hypoxia-inducible factor 1	Yes
SP1	SP1 transcription factor	Yes
GLI	GLI-Kruppel family member	Yes
EGR3	Early growth response 3	Yes

PPAR, and *SP1* have high influence via DTI and are expressed in embryonic kidney (Table 4), apart from having conserved TFBSs. This lends good computational evidence for the role of these TFs in *Gata3* regulation, and presents a reasonable hypothesis worthy of experimental validation.

Additionally, we examined the influence for another two TFs, *STE12* and *HP1*, both of which have a high coexpression correlation with *Gata3* as well as conserved TFBSs in the promoter region. The DTI criterion gave us no evidence of influence between these two TFs and *Gata3* activity. This information, coupled with the present evidence concerning the nonkidney specificity of *STE12* and *HP1*, presents an argument for the noninvolvement of these TFs in kidney-specific regulation of *Gata3*. Thus, the DTI criterion can be used to guide more focused experiments to identify the true transcriptional effectors underlying *Gata3* expression.

This application shows the utility of DTI to specifically explore the expression-level influence of a TF of interest to any target gene. This result, coupled with the unsupervised network inference methods in kidney and T-cell data, establishes the DTI-based methodology as a common framework for both types of analysis.

8.5. *Module TFs in coregulated genes*

We examine another interesting scenario for the principled application of the DTI criterion. The coexpression of genes in a biological context is a complex phenomenon involving the combinatorial regulation of such genes by several TFs. Such coexpression occurs during processes like development and disease progression. This is also observed in coclustered genes from the output of hierarchical clustering algorithms (signatures). The underlying hypothesis is that coclustered/coexpressed genes might be under the control of some common TFs (modules), which underlie the coordinated expression of all these implicated genes.

Several tools (e.g. Genomatix,[38] CREME,[39] Toucan[40]) allow the inference of such TF modules from sets of genes. However, the next logical question is whether any of the TFs comprising the module indeed have an expression-level influence on these target gene(s). Supervised DTI can be used in this context to rank the most likely effector TFs for each gene in the gene set.

To illustrate this application, genes that have expression in the developing ureteric bud (UB) in the kidney are examined. The inductive signals between the UB and MM cause the differentiation of fetal kidney stem cells into nephrons, the basic unit of function of the kidney. An examination of these UB-specific genes (obtained from the Mouse Genome Informatics (MGI) repository at http://www.informatics.jax.org/)[16,41] reveals some modules. The UB-specific genes and the modules are listed in Tables 5 and 6, respectively.

Briefly, the modules are obtained as follows. The various UB-specific gene sequences are mined for their proximal promoter (from ~2,000 bp upstream to

Table 5. Genes expressed in the developing ureteric bud (days e10.5–11.0), as reported in the Mouse Genome Informatics database.

Ensembl gene ID	Gene name
ENSMUSG00000015619	*Gata3*
ENSMUSG00000032796	*Lama1*
ENSMUSG00000015647	*Lama5*
ENSMUSG00000026478	*Lamc1*
ENSMUSG00000018698	*Lhx1*
ENSMUSG00000008999	*Bmp7*
ENSMUSG00000023906	*Cldn6*
ENSMUSG00000059040	*Eno1*
ENSMUSG00000004231	*Pax2*
ENSMUSG00000030110	*Ret*
ENSMUSG00000022144	*Gdnf*
ENSMUSG00000031681	*Smad1*
ENSMUSG00000024563	*Smad2*
ENSMUSG00000074227	*Spint2*
ENSMUSG00000015957	*Wnt11*
ENSMUSG00000039481	*Nrtn*
ENSMUSG00000063358	*Mapk1*
ENSMUSG00000063065	*Mapk3*

Table 6. Annotation of the module TFs from UB-specific genes.

TFs in module	Annotation	Kidney specificity (Yes/No) (GNF/literature)
SP1	*trans*-acting TF 1	Y
LMO2	LIM domain only 2	N
OCT1	POU domain, class 2, TF 1	Y
ZIC1	Zinc finger protein of the cerebellum 1	N
MZF1	Myeloid zinc finger 1	Y
AP2	TF AP-2	Y
AP4	TF AP-4	Y
YY1	YY1 transcription factor	Y
TAL1	T-cell acute lymphocytic leukemia 1	Y (cell line)
PAX2	Paired box gene 2	Y
HNF4	Hepatocyte nuclear factor 4	Y

GNF: Genomics Institute of the Novartis Research Foundation.

200 bp downstream from the transcription start site). The various promoters are then aligned, and a search for significantly overrepresented TFs is done using the position-weight matrices derived from the TRANSFAC/JASPAR database (MotifScanner). From this set of TFs, modules of TFs (with potentially overlapping sites) are obtained (ModuleSearcher). The TOUCAN 3.0.2 tool[40] allows for the entire sequence of steps from sequence extraction to module searches. All TFs in the various modules identified are listed in Table 6.

The list of module TFs is obtained by combining expression annotations (from MGI) and sequence analysis. For the purpose of integrating heterogeneous data and to reduce the number of candidate TFs that are putatively involved in regulating UB-specific genes, we can use DTI to find influences between the TF genes and the UB-specific genes using expression data. As an example, one of the TFs in the module list is *Pax2* and it has an important role in UB differentiation[33]; another gene expressed in the developing UB is *Gata3*. We now examine if the DTI $I(Pax2 \rightarrow Gata3)$ is significant and ranks high in the list. This is highlighted in Fig. 8.

For the *Pax2–Gata3* interaction, we show the cumulative distribution function of the bootstrapped detection statistic (Fig. 8) as well as the position of the true DTI estimate in relation to the overall histogram. With the obtained density estimate of the *Pax2–Gata3* interaction, in Fig. 8, we can find significance values of the true DTI estimate in relation to the bootstrapped null distribution.

An experimental validation of this is presented in Grote *et al.*[33] and Dressler and Douglas.[42] Thus, we can look at each module member for a possible role in *Gata3* regulation. As can be seen, this approach integrates sequence information, phylogeny, and expression to look for upstream effectors for genes of interest (i.e. those that share some pattern of coexpression/coregulation).

Extending this further, the strength and significance of the DTI can be found between every pair of TF and UB-specific genes of Tables 5 and 6. This can be visualized as a bipartite graph of TF–gene interactions, shown in Fig. 9. The graph summarizes the degree of interactions between the various TFs in the modules

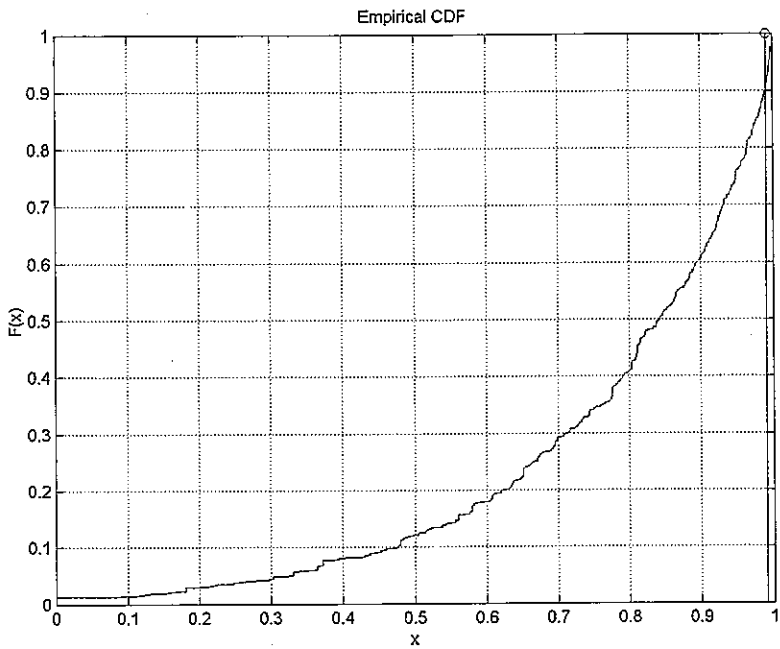

Fig. 8. Cumulative distribution function for bootstrapped $I(Pax2 \rightarrow Gata3)$. The true value of $I(Pax2 \rightarrow Gata3) = 0.9911$.

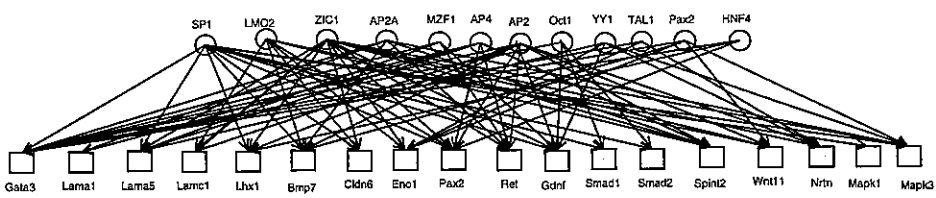

Fig. 9. A bipartite graph between the group of module TFs and genes coexpressed in the developing ureteric bud (MGI: e10.5–11.0).

and the coexpressed genes; and is the overall integration of annotation, sequence, and expression data. Additionally, the embryonic kidney specificity of the various module TFs is listed, based on literature and tissue-specificity annotation (http://symatlas.gnf.org/SymAtlas/). It is to be noted that some TFs such as *SP1* have ubiquitous expression across most tissues,[43,44] and are not as informative as kidney-specific ones like *Pax2*[33] or *HNF4a*.[45]

8.6. *Higher-order MI and DTI*

The final part of this work highlights that directed information (DTI) and mutual information (MI) can together aid in the discovery of higher-order interactions among genes. Higher-order MI[9,19] has been used successfully for the discovery of

interactions among triples of genes. Following work done in Schneidman *et al.*,[46] we use the triplet information given by

$$\begin{aligned} I_3(x_i;x_j;x_k) &= \sum_i H(x_i) - \sum_{i<j} H(x_i,x_j) + H(x_i,x_j,x_k) \\ &= I(x_1;x_2;x_3) - \sum_{i<j} I(x_i;x_j) \\ &= [I(x_1;x_3) + I(x_2;x_3)] - I(\{x_1,x_2\};x_3). \end{aligned}$$

From the above definition, it is clear that the use of triplet information helps resolve the pairwise-joint dependencies between x_i, x_j, and x_k versus the synergistic dependence of any variable on the combination of the other two variables. A positive value of $I_3(x_i;x_j;x_k)$ indicates pairwise dependence, and thus DTI can be used to infer directional association between x_i, x_j, and x_k. A negative value indicates synergy and needs to be resolved further.

For the network shown in Fig. 5, we aim to recover any synergistic interactions of various genes, using higher-order entropy methods, that are potentially missed due to consideration of only pairwise interactions.

For the synergy framework presented above, we seek to determine the direction of association of $\{x_i, x_j\}$ and x_k, for all genes i, j, k. For this purpose, $I(\{x_i, x_j\} \rightarrow x_k)$ is determined, using methods presented earlier. Depending on the relative magnitude of $I(\{x_i, x_j\} \rightarrow x_k)$ and $I(x_k \rightarrow \{x_i, x_j\})$, the direction of association can be determined.

We now consider the set of genes common to those profiled in the microarray expression[15,16,47] study as well as the annotated genes from MGI. For these 12 genes (*Bmp7*, *Cldn7*, *Gata3*, *Gdnf*, *Lamc2*, *Mapk1*, *Mapk3*, *Nrtn*, *Pax2*, *Ret*, *Spint1*, *Wnt11*), we study the dependencies discovered using triplet information. Also, for the purposes of this work, we only present those dependencies wherein the triplet information is negative, indicating possible synergistic interactions. These interactions are indicated below (Table 7).

Several of the pathways, such as *Gdnf–Ret*, *Wnt*, and *Mapk*, are implicated in UB differentiation.[48,49] However, most studies have focused on interaction within

Table 7. Some triplet interactions (discovered using DTI) that have a putative biological role. Biological validation from the literature is given in parentheses.

			UB specificity and citation (http://symatlas.gnf.org/SymAtlas/)
Gdnf	*Ret*	*Gata3*	Y (Grote *et al.*[33])
Ret	*Bmp7*	*Gata3*	Y (Davies[49])
Pax2	*Gata3*	*Ret*	Y (Clarke *et al.*[50])
Ret	*Wnt11*	*Gdnf*	Y (Majumdar *et al.*[48])
Pax2	*Wnt11*	*Gata3*	Y (Grote *et al.*[33])
Pax2	*Ret*	*Gdnf*	Y (Clarke *et al.*[50] and Brophy *et al.*[51])

a pathway and not so much on cross-talk between various pathways. Organ development is a complex phenomenon and needs several reciprocal interactions to control the growth of various cell populations. It is interesting to see several known cross-interactions picked up using higher-order information, based on expression data alone (Table 7). Given that cooperation/synergies between various pathways are essential in most other biological processes, we believe that using a combination of higher-order MI and DTI would aid in the experimental resolution of such interactions.

9. Conclusions

In this work, we have presented the notion of directed information (DTI) as a reliable criterion for the inference of influence in gene networks. After motivating the utility of DTI in discovering directed nonlinear interactions, we present two variants of DTI that can be used, depending on the context. One version, unsupervised DTI, like traditional network inference, enables the discovery of influences (regulatory or nonregulatory) among any given set of genes; the other version, supervised DTI, aids the modeling of the strength of influence between two specific genes of interest — questions arising during transcriptional influence. It is interesting that DTI enables the use of a common framework for both these purposes and is general enough to accommodate arbitrary lag, nonlinearity, and resolution of cycles, loops, and direction.

We see that the above presented combination of supervised and unsupervised variants enables their applicability to several important problems in bioinformatics (e.g. upstream TF discovery, module–gene interactions, higher-order influence determination), some of which are presented in Sec. 8. The network inference approach can also allow incorporation of additional biophysical knowledge pertaining to both physical mechanisms and protein interactions that exist during transcription. We point out that, given the diverse nature of biological data of varying throughput, one has to adopt an approach to integrate such data to make biologically relevant findings. Hence, the DTI metric fits very naturally into such an integrative framework.

Acknowledgments

The authors gratefully acknowledge the support of the NIH under award 5R01-GM028896-21 (J. D. E.) We would like to thank Prof. Sandeep Pradhan and Mr. Ramji Venkataramanan for useful discussions on directed information. We are very grateful to Prof. Erik Learned-Miller for sharing his code for higher-order entropy estimation, and Prof. Bruce Aronow for kidney expression data. We are also grateful to the reviewers for carefully reading and offering several helpful insights to improve the quality of the manuscript.

References

1. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A, ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics* **7**(Suppl 1):S7, 2006.
2. Opgen-Rhein R, Strimmer K, Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data, in *Proc Fourth International Workshop on Computational Systems Biology (WCSB)*, pp. 73–76, 2006.
3. Woolf PJ, Prudhomme W, Daheron L, Daley GQ, Lauffenburger DA, Bayesian analysis of signaling networks governing embryonic stem cell fate decisions, *Bioinformatics* **21**(6):741–753, 2005.
4. Marko H, The bidirectional communication theory — A generalization of information theory, *IEEE Trans Commun* **21**:1345–1351, 1973.
5. Massey J, Causality, feedback and directed information, in *Proc 1990 Symp Information Theory and Its Applications (ISITA-90)*, Waikiki, HI, pp. 303–305, 1990.
6. Venkataramanan R, Pradhan SS, Source coding with feed-forward: Rate-distortion theorems and error exponents for a general source, *IEEE Trans Inf Theory* **53**(6):2154–2179, 2007.
7. Geweke J, The measurement of linear dependence and feedback between multiple time series, *J Am Stat Assoc* **77**:304–324, 1982. (With comments by Parzen E, Pierce DA, Wei W, Zellner A, and rejoinder)
8. Rao A, Hero AO, States DJ, Engel JD, Inference of biologically relevant gene influence networks using the directed information criterion, in *Proc IEEE Conf Acoustics, Speech and Signal Processing*, pp. 1028–1031, 2006.
9. Nemenman I, Information theory, multivariate dependence, and genetic network inference, Technical Report NSF-KITP-04-54, Kavli Institute for Theoretical Physics, University of California at Santa Barbara, Santa Barbara, CA, 2004.
10. Khandekar M, Suzuki N, Lewton J, Yamamoto M, Engel JD, Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system, *Mol Cell Biol* **24**(23):10263–10276, 2004.
11. MacIsaac KD, Fraenkel E, Practical strategies for discovering regulatory DNA sequence motifs, *PLoS Comput Biol* **2**(4):e36, 2006.
12. Kreiman G, Identification of sparsely distributed clusters of *cis*-regulatory elements in sets of co-expressed genes, *Nucleic Acids Res* **32**(9):2889–2900, 2004.
13. Lakshmanan G, Lieuw KH, Lim KC, Gu Y, Grosveld F, Engel JD, Karis A, Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus, *Mol Cell Biol* **19**(2):1558–1568, 1999.
14. Ovcharenko I, Nobrega MA, Loots GG, Stubbs L, ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes, *Nucleic Acids Res* **32**:W280–W286, 2004.
15. Challen G, Gardiner B, Caruana G, Kostoulias X, Martinez G, Crowe M, Taylor DF, Bertram J, Little M, Grimmond SM, Temporal and spatial transcriptional programs in murine kidney development, *Physiol Genomics* **23**(2):159–171, 2005.
16. Schwab K, Patterson LT, Aronow BJ, Luckas R, Liang HC, Potter SS, A catalogue of gene expression in the developing kidney, *Kidney Int* **64**(5):1588–1604, 2003.
17. Cover TM, Thomas JA, *Elements of Information Theory*, Wiley-Interscience, New York, 1991.
18. Miller E, Miller E, A new class of entropy estimators for multi-dimensional densities, in *Proc IEEE Int Conf Acoustics, Speech, and Signal Processing (ICASSP)*, III-297–III-300, 2003.

19. Nemenman I, Shafee F, Bialek W, Entropy and inference, revisited, in Dietterich TG, Becker S, Ghahramani Z (eds.), *Advances in Neural Information Processing Systems*, Vol. 14, MIT Press, Cambridge, MA, 2002.
20. Paninski L, Estimation of entropy and mutual information, *Neural Comput* **15**:1191–1254, 2003.
21. Willett R, Nowak R, Complexity-regularized multiresolution density estimation, in *Proc Int Symp Information Theory (ISIT)*, p. 305, 2004.
22. Learned-Miller E, Hyperspacings and the estimation of information theoretic quantities, Technical Report 04-104, University of Massachusetts Amherst, Amherst, MA, 2004.
23. Effron B, Tibshirani RJ, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York, 1994.
24. Ramsay J, Silverman BW, *Functional Data Analysis*, Springer, New York, 1997.
25. Moonen CTW, Bandettini PA (eds.), *Functional MRI*, Springer, Berlin, 2000.
26. Benjamini Y, Hochberg Y, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J R Stat Soc Ser B* **57**:289–300, 1995.
27. Schäfer J, Strimmer K, An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics* **21**:754–764, 2005.
28. Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, Gaiba A, Wild DL, Falciani F, Modeling T-cell activation using gene expression profiling and state-space models, *Bioinformatics* **20**(9):1361–1372, 2004.
29. Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL, A Bayesian approach to reconstructing genetic regulatory networks with hidden factors, *Bioinformatics* **21**(3):349–356, 2005.
30. Hashimoto RF, Kim S, Shmulevich I, Zhang W, Bittner ML, Dougherty ER, Growing genetic regulatory networks from seed genes, *Bioinformatics* **20**(8):1241–1247, 2004.
31. Li H, Sun Y, Zhan M, Analysis of gene coexpression by B-spline based CoD estimation, *EURASIP J Bioinform Syst Biol* **2007**:6, 2007.
32. Balmer JE, Blomhoff R, Gene expression regulation by retinoic acid, *J Lipid Res* **43**(11):1773–1808, 2002.
33. Grote D, Souabni A, Busslinger M, Bouchard M, Pax 2/8-regulated Gata3 expression is necessary for morphogenesis and guidance of the nephric duct in the developing kidney, *Development* **133**(1):53–61, 2006.
34. Zhang SL, Moini B, Ingelfinger JR, Angiotensin II increases Pax-2 expression in fetal kidney cells via the AT2 receptor, *J Am Soc Nephrol* **15**(6):1452–1465, 2004.
35. Ezzat S, Mader R, Yu S, Ning T, Poussier P, Asa SL, Ikaros integrates endocrine and immune system development, *J Clin Invest* **115**(4):844–848, 2005.
36. Zhang DH, Yang L, Ray A, Differential responsiveness of the IL-5 and IL-4 genes to transcription factor GATA-3, *J Immunol* **161**:3817–3821, 1998.
37. Rogoff HA, Pickering MT, Frame FM, Debatis ME, Sanchez Y, Jones S, Kowalik TF, Apoptosis associated with deregulated E2F activity is dependent on E2F1 and Atm/Nbs1/Chk2, *Mol Cell Biol* **24**(7):2968–2977, 2004.
38. Cohen CD, Klingenhoff A, Boucherot A, Nitsche A, Henger A, Brunner B, Schmid H, Merkle M, Saleem MA, Koller KP, Werner T, Grone HJ, Nelson PJ, Kretzler M, Comparative promoter analysis allows *de novo* identification of specialized cell junction-associated proteins, *Proc Natl Acad Sci USA* **103**(15):5682–5687, 2006.
39. Papatsenko D, Levine M, Computational identification of regulatory DNAs underlying animal development, *Nat Methods* **2**:529–534, 2005.

40. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B, Toucan: Deciphering the *cis*-regulatory logic of coregulated genes, *Nucleic Acids Res* **31**(6):1753–1764, 2003.
41. Stuart RO, Bush KT, Nigam SK, Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development, *Kidney Int* **64**(6):1997–2008, 2003.
42. Dressler GR, Douglas EC, Pax-2 is a DNA-binding protein expressed in embryonic kidney and Wilms tumor, *Proc Natl Acad Sci USA* **89**:1179–1183, 1992.
43. Cohen HT, Bossone SA, Zhu G, McDonald GA, Sukhatme VP, Sp1 is a critical regulator of the Wilms' tumor-1 gene, *J Biol Chem* **272**(5):2901–2913, 1997.
44. Ryan G, Steele-Perkins V, Morris JF, Rauscher FJ, Dressler GR, Repression of Pax-2 by WT1 during normal kidney development, *Development* **121**(3):867–875, 1995.
45. Taraviras S, Monaghan AP, Schtz G, Kelsey G, Characterization of the mouse HNF-4 gene and its expression during mouse embryogenesis, *Mech Dev* **48**(2):67–79, 1994.
46. Schneidman E, Still S, Berry MJ II, Bialek W, Network information and connected correlations, *Phys Rev Lett* **91**:238701, 2003.
47. Challen GA, Martinez G, Davis MJ, Taylor DF, Crowe M, Teasdale RD, Grimmond SM, Little MH, Identifying the molecular phenotype of renal progenitor cells, *J Am Soc Nephrol* **15**(9):2344–2357, 2004.
48. Majumdar A, Vainio S, Kispert A, McMahon J, McMahon AP, Wnt11 and Ret/Gdnf pathways cooperate in regulating ureteric branching during metanephric kidney development, *Development* **130**(14):3175–3185, 2003.
49. Davies J, Intracellular and extracellular regulation of ureteric bud morphogenesis, *J Anat* **198**(3):257–264, 2001.
50. Clarke JC, Patel SR, Raymond RM, Andrew S, Robinson BG, Dressler GR, Brophy PD, Regulation of c-Ret in the developing kidney is responsive to Pax2 gene dosage, *Hum Mol Genet* **15**(23):3420–3428, 2006.
51. Brophy PD, Ostrom L, Lang KM, Dressler GR, Regulation of ureteric bud outgrowth by Pax2-dependent activation of the glial derived neurotrophic factor gene, *Development* **128**(23):4747–4756, 2001.
52. Gubner JA, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, Cambridge, UK, 2006.
53. Joe H, Relative entropy measures of multivariate dependence, *J Am Stat Assoc* 84:157–164, 1989.

Appendix: A Normalized DTI Measure

In this section, an expression for a normalized DTI coefficient is derived. This is useful for a meaningful comparison across different criteria during network inference. The purpose of this section is to establish some connections between quantities like MI, DTI, and correlation. In this section, we use X, Y, and Z for X^N, Y^N, and Z^N interchangeably, i.e $X \equiv X^N$, $Y \equiv Y^N$, and $Z \equiv Z^N$.

By the definition of DTI, we can see that $0 \leq I(X^N \to Y^N) \leq I(X^N; Y^N) < \infty$. The normalized measure ρ_{DTI} should be able to map this large range ($[0, \infty]$) to $[0, 1]$. We recall that the multivariate canonical correlation is given by[52] $\rho_{X^N;Y^N} = \Sigma_{X^N}^{-1/2} \Sigma_{X^N Y^N} \Sigma_{Y^N}^{-1/2}$, and this is normalized having eigenvalues between 0 and 1. We also recall that, under a Gaussian distribution on X^N and Y^N, the joint entropy $H(X^N; Y^N) = -\frac{1}{2} \ln(2\pi e)^{2N} |\Sigma_{X^N Y^N}|$, where $|\Sigma_{X^N Y^N}|$ is the determinant of matrix $\Sigma_{X^N Y^N}$, and $\Sigma_{X^N Y^N}$ denotes the covariance matrix, computed as

$\Sigma_{X^N Y^N} = \frac{1}{R-1} X^T Y$, indicating that there are R replicates of the X, Y time series, each of length N.

Thus, for $I(X^N;Y^N) = H(X^N) + H(Y^N) - H(X^N,Y^N)$, the expression for mutual information, under jointly Gaussian assumptions on X^N and Y^N, becomes $I(X;Y) = -\frac{1}{2}\ln\left(\frac{|\Sigma_{X^N Y^N}|^2}{|\Sigma_{X^N}|\cdot|\Sigma_{Y^N}|}\right) = -\frac{1}{2}\ln(1-\rho^2_{X^N;Y^N})$. Hence, a straightforward transformation is normalized MI: $\rho_{\mathrm{MI}} = \sqrt{1-e^{-2I(X^N;Y^N)}} = \sqrt{1-e^{-2\sum_{i=1}^N I(X^N;Y_i|Y^{i-1})}}$. A connection with Joe[53] can thus be immediately seen.

With this, ρ_{MI} is normalized between $[0,1]$ and gives a better absolute definition of dependency that does not depend on the unnormalized MI. We will use this definition of normalized information coefficients in the present set of simulation studies.

For constructing a normalized version of the DTI, we can extend the approach from Geweke.[7] Consider three random vectors $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, each of which is identically distributed as $\mathcal{N}(\mu_X, \Sigma_{XX})$, $\mathcal{N}(\mu_Y, \Sigma_{YY})$, and $\mathcal{N}(\mu_Z, \Sigma_{ZZ})$, respectively. We also have

$$(\mathbf{X},\mathbf{Y},\mathbf{Z}) \sim \mathcal{N}\left[\begin{pmatrix}\mu_X\\ \mu_Y \mu_Z\end{pmatrix}, \begin{pmatrix}\Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ}\\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ}\\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ}\end{pmatrix}\right].$$

Their partial correlation $\delta_{YX|Z}$ is then given by $\delta_{YX|Z} = \sqrt{\frac{a_2^2}{a_1 a_3}}$, where $a_1 = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$, $a_2 = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$, and $a_3 = \Sigma_{XX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$.

Recalling results from conditional Gaussian distributions, these can be denoted by $a_1 = \Sigma_{Y|Z}$, $a_2 = \Sigma_{XY|Z}$, and $a_3 = \Sigma_{X|Z}$. Thus, $\delta_{YX|Z} = \Sigma_{Y|Z}^{-1/2}\Sigma_{XY|Z}\Sigma_{X|Z}^{-1/2}$. Extending the above result from the MI to the DTI case, we have $\rho_{\mathrm{DTI}} = \sqrt{1-e^{-2\sum_{i=1}^N I(X^i;Y_i|Y^{i-1})}}$.

We recall the primary difference between MI and DTI (note the superscript on X):

$$\text{MI}: \quad I(X^N;Y^N) = \sum_{i=1}^N I(X^N;Y_i|Y^{i-1}).$$

$$\text{DTI}: \quad I(X^N \to Y^N) = \sum_{i=1}^N I(X^i;Y_i|Y^{i-1}).$$

Having found the normalized DTI, we ask if the obtained DTI estimate is significant with respect to a null DTI distribution obtained by random chance. This is addressed in Sec. 6.

We note that, though the normality assumption was used to show the connection between information and correlation, this distributional assumption is not used anywhere in the original DTI metric formulation and computation during its application to network inference.

Arvind Rao received his B.E. from Bangalore University, India (2001); his M.S.E. (Communications, Networks and Systems) from the University of Texas at Austin, USA (2003); and his M.A. in Statistics (2007) from the University of Michigan, USA. He is currently a joint Ph.D. candidate in the Departments of Electrical Engineering and Computer Science (Systems Division) and the Bioinformatics Graduate Program at the University of Michigan. His current research interests are in systems biology and machine learning in the context of understanding genome regulation. At Michigan, he is currently a Rackham Predoctoral Fellow. He is a member of the IEEE Signal Processing Society, SIAM, and the Life Sciences Society.

Alfred O. Hero III received his B.S. (*summa cum laude*) from Boston University, USA (1980) and his Ph.D. from Princeton University, USA (1984), both in Electrical Engineering. Since 1984, he has been with the University of Michigan, USA, where he is a Professor in the Department of Electrical Engineering and Computer Science and, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. His recent research interests have been in areas including inference in sensor networks, adaptive sensing, bioinformatics, inverse problems, and statistical signal and image processing. He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), and has received an IEEE Signal Processing Society Meritorious Service Award (1998), an IEEE Signal Processing Society Best Paper Award (1998), and the IEEE Third Millenium Medal (2000). He was President of the IEEE Signal Processing Society from 2006 to 2007.

David J. States received his B.A. (*magna cum laude*) from Harvard College, USA (1975) as well as his M.D. and Ph.D from Harvard University (1983) in Biophysics. He was a resident in Internal Medicine at the University of California, San Diego, and a Clinical Associate at the National Heart Lung and Blood Institute. In 1988, he joined the National Center for Biotechnology Information, and in 1992 moved to Washington University in St. Louis to be the Director of the Institute for Biomedical Computing. Since 2001, he has been with the University of Michigan, where he is a Professor in the Department of Human Genetics. His recent research interests have been in areas including alternative splicing in cancer, the analysis of genome regulation, and data integration in molecular biology. He is a Fellow of the American College of Medical Informatics (ACMI). He was a founding member of the Board of Directors of the International Society for Computational Biology, and served as Treasurer from 1997 to 2000. He chaired two Organizing Committees for the Intelligent Systems in Molecular Biology conference in 1996 and 2005.

James Douglas Engel received his Ph.D. in Biophysical Chemistry at the University of Oregon, USA, in 1975 with P. H. von Hippel. He was a Helen Hay Whitney Fellow with N. Davidson and T. Maniatis at Caltech from 1975 to 1978, and then joined the faculty at Northwestern University, where he became the Owen L. Coon Professor and Associate Director for Basic Sciences of the Robert H. Lurie Comprehensive Cancer Center. He moved to the University of Michigan School of Medicine

in 2002, and was endowed the first G. Carl Huber Chair in Developmental Biology and Chair of the Department of Cell and Developmental Biology. Dr. Engel's lab is devoted to deciphering, both experimentally and theoretically, the transcriptional regulatory networks that lead to correctly modulated gene expression during embryonic development. Dr. Engel is a Fellow of the American Association of Arts and Sciences, and an editor of *Molecular and Cellular Biology.*