USING DIRECTED INFORMATION TO BUILD BIOLOGICALLY RELEVANT INFLUENCE NETWORKS

Arvind Rao* and Alfred O. Hero, III
Electrical Engineering and Computer Science, Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
**Email: [ukarvind, hero]@umich.edu*

David J. States
Bioinformatics, Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
Email: dstates@umich.edu

James Douglas Engel
Cell and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA
Email: engel@umich.edu

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, which 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 as well as T-cell activation. Conformity of the obtained interactions with literature as well as comparison with the coefficient of determination (CoD) method is 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 between transcription factor modules and target co-regulated genes is proposed. Additionally, we show that DTI can be used in conjunction with mutual information to infer higher-order influence networks involving co-operative gene interactions.

Keywords: Mutual Information; Directed Information; transcription factor module; comparative genomics; transcription regulatory network.

1. INTRODUCTION

Computational methods for inferring dependencies between genes [31, 36, 52] 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). The DTI is a variant of mutual information (MI) that attempts to capture the di-

*Corresponding author.

rection of information flow. It is widely used in the analysis of communication systems with feedback or feedforward [32, 33, 50] as well as in economic time series analysis [17, 50]. The DTI [32, 42] can be interpreted as a directed version of mutual information, a criterion used quite frequently in other related work [31]. It turns out, as we will demonstrate, that the 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 co-expressed genes. Higher-order network inference, in this work, examines the resolution of three-way interactions rather than only pairwise relationships among genes [35].

2. ORGANIZATION

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

3. GENE NETWORKS

Transcription is the process of generation of messenger RNA (mRNA) from the DNA template representing the gene. It is the intermediate step before the generation of functional protein from messenger RNA. During gene expression (Fig. 1), transcription factor 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 [23]. Since transcription factors are also proteins (or their activated forms) which are in turn encoded for by other genes, the notion of an influence between a transcription factor gene and the target gene can be considered.

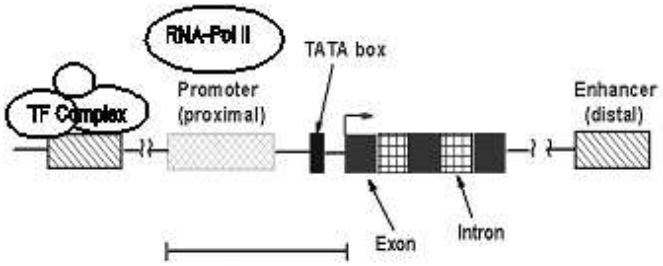


Fig. 1. Schematic of Transcriptional Regulation. Sequence motifs at the promoter and the distal regulatory elements together confer specificity of gene expression via TF binding.

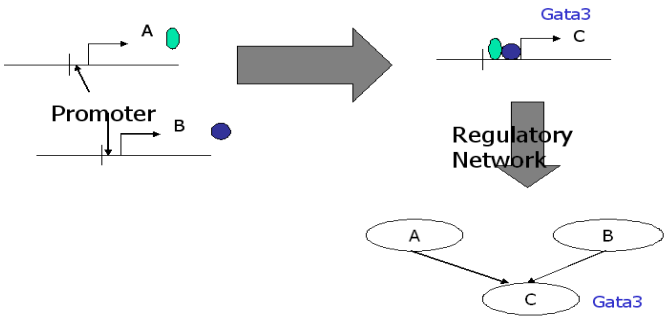


Fig. 2. A transcriptional regulatory network with genes A and B effect C. An example of C that we study here is the *Gata3* gene.

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 pro-

tein 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 transcription factor (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 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 [29, 24].

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 [25, 23], like in the developing urogenital system (nephrogenesis), central nervous system, and T-cell development. In order to find which TFs regulate the tissue-specific transcription of *Gata3* (either at the promoter or long-range regulatory elements), a commonly followed approach [24, 29] is to look for phylogenetically conserved transcription factor binding sites (TFBS). 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 questions:

- Biologists are also interested in the network of relationships among genes expressed under a certain set of conditions, which uses several network inference procedures, such as Bayesian networks [52], mutual information [31] etc. However, there has been lack of a common framework to do both supervised *and* unsupervised *directed* network inference within these settings to detect directed non-linear gene-gene interactions. We present directed information 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 sections 8.2 and 8.3.
- Which transcription factors are potentially active at the target gene's promoter during its tissue-specific regulation ? - this question is primarily answered by examining the phylogenetically conserved TFBS at the promoter and asking if microarray expression data suggests 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 (section: 8.4).
- Which transcription factors underlie the tissue-specific expression of a group of co-expressed/co-regulated genes (eg: *Gata3* and others)? - one common approach is to search the proximal promoters of all such tissue specific genes, and look for 'modules' of TFs that control tissue-specific expression [24, 29]. 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 co-expressed gene promoters and use microarray expression to point out possible effectors of target gene expression (section: 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.

Towards this end, the inference of higher order interactions (more than only two-way gene relationships) is seen to be a useful approach [35]. Using DTI, it would be interesting to find directed interactions between differentially expressed genes of the developing kidney to determine pathway cross-talk (section: 8.6).

4.1. Phylogenetic Conservation of Transcription Factor Binding Sites (TFBS)

As mentioned above, the mechanism of regulation of a target gene is via the binding site of the corresponding transcription factor (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, Chimp, Mouse, Rat and Chicken), then we can look for the conservation of functional binding sites over all 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 TFBS (Fig. 3). Such information can be obtained from most genome browsers [37]. We see that even for a fairly short stretch of sequence (1 kilobase) upstream of the gene, there are several conserved sequence elements which are potential TFBS (light grey regions in Fig. 3).

In this figure, 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 transcription factors 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 contain 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), metanephric mesenchyme (MM) apart from the whole kidney.

Our approach thus integrates several aspects:

- Using phylogenetic information to infer which TF binding sites upstream of a target gene may be functional.
- Identifying if any of the TF genes influence a target gene by coding for a transcription factor that binds at the site discovered from conservation studies. This directed influence is captured using an influence metric (like directed information) in conjunction with expression data ([8, 47])and explained in Section: 5.

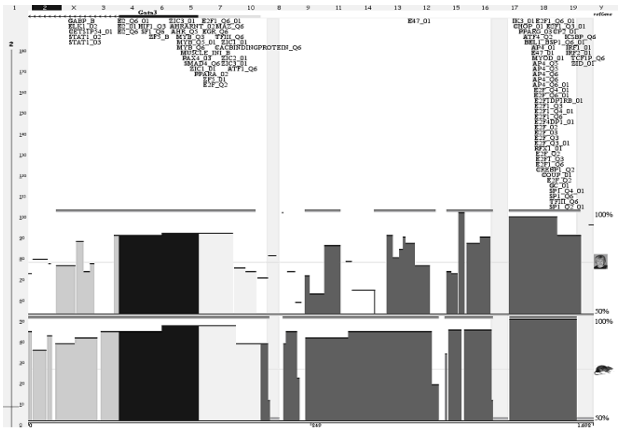


Fig. 3. TFBS conservation between Human, Mouse and Rat, upstream (x-axis) of *Gata3*, from *http://www.ecrbrowser.dcode.org/*.

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 (like *Gata3*). Several such metrics have been proposed, notably, correlation, coefficient of determination (CoD), mutual information etc. To alleviate the challenge of detecting non-linear 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 [31]. 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 criterion (DTI), for the inference of non-linear, directed gene influences.

The 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 [33]:

$$I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1}) \qquad (1)$$

Here, Y^n denotes $(Y_1, Y_2, .., Y_n)$, i.e. a segment of the realization of a random process Y and $I(X^N; Y^N)$ is the Shannon mutual information [12].

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, the 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 and due to computational complexity and moderate sample size, histogram estimation of multivariate density is unviable. However, several methods exist for consistent entropy estimation of multivariate small sample data [26, 34, 38, 51]. In the context of microarray expression data, wherein probe-level and technical/biological replicates are available, we use the method of [26] for entropy estimation.

From (1), we have,

$$\begin{aligned} I(X^N \rightarrow Y^N) &= \sum_{n=1}^{N} [H(X^n|Y^{n-1}) - H(X^n|Y^n)] \\ &= \sum_{n=1}^{N} \{[H(X^n, Y^{n-1}) - H(Y^{n-1})] - \\ &\qquad [H(X^n, Y^n) - H(Y^n)]\} \end{aligned} \qquad (2)$$

- To evaluate the DTI expression in *Eqn.*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 are R dimensional, R being the number of replicates (probe-level, biological and technical).
- Though 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 the use the method proposed in [26]. 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 k^{th} and l^{th} time instants of gene expression.
- 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 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 [27].

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 the above expression

(2).

From the definition of DTI, we know that $0 \leq I(X_i^N \rightarrow Y^N) \leq I(X_i^N; Y^N) < \infty$. For easy comparison with other metrics, we use a normalized DTI metric (see Appendix) given by $\rho_{DI} = \sqrt{1 - e^{-2I(X^N \rightarrow Y^N)}} = \sqrt{1 - e^{-2\sum_{i=1}^{N} I(X^i; Y_i | Y^{i-1})}}$. This maps the large range of DI, ($[0, \infty]$) 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 chance 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 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 to ascertain the significance of the normalized directed information $\hat{I}(X^N \rightarrow Y^N)$ 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 $\Theta = \hat{I}(X^N \rightarrow Y^N)$ and the chosen acceptable p-value is α.

The overall bootstrap based test procedure is ([15],[40],[2]):

- Repeat the following procedure $B(= 1000)$ times (with index $b = 1, \ldots, B$):
  - Generate resampled (with replacement) versions of the times series X^N, Y^N, denoted by X_b^N, Y_b^N respectively.
  - Compute the statistic $\theta^b = \hat{I}(X_b^N \rightarrow Y_b^N)$.
- Construct an empirical CDF (cumulative distribution function) 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) $\theta_0 = \hat{I}(X^N \rightarrow Y^N)$ 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 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 characteristic 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 (Eqn. 2).
- For each pair of genes A_i and B among these G genes :
  {
  - Look for a phylogenetically conserved binding site of TF encoded by gene A_i in the upstream region of gene B.
  - Find $DTI(A_i, B) = I(A_i^N \rightarrow B^N)$, and the normalized DTI from A_i to B, $DTI(A_i, B) = \sqrt{1 - e^{-2I(A_i^N \rightarrow B^N)}}$.
  - 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 $((A_1, A_2, \ldots, A_G))$.

  }

Note: 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 [5] is used to screen each of the $M(= G(G-1))$ hypotheses (both directions) during network discovery amongst 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 [45]. 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 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), 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 non-linear interactions. This is because interactions amongst genes involve non-linear kinetics.
- Be a non-parametric procedure to avoid distributional assumptions (noise etc).
- Be capable of recovering interactions that a biologist might be interested in. Rather than use a method that discovers interactions underlying the data purely, the biologist should be able to use prior knowledge (from 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 these considerations.

Table 1. Comparison of various network inference methods.

Method	Resolve Cycles	Non -linear framework	Search for interaction	Non -parametric framework
SSM [41, 4]	Y	Y	N	Y
CoD [20]	N	N	Y	N
GGM [36]	N	Y	N	N
DTI [42]	Y	Y	Y	Y

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 transcription factors 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 TFBSes, 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 development. The idea is to look for common transcription factor modules that govern the combined co-expression and co-regulation of these genes [29]. 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.

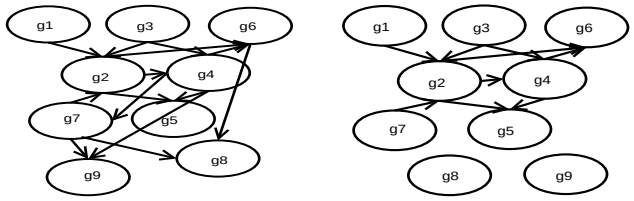
Before proceeding, we examine the performance of this approach on synthetic data.

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 Coefficient of Determination (CoD) approach for directed influence network determination. The CoD allows the determination of association between two genes via a R^2 goodness of fit statistic. The methods of [20, 28] 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 use some underlying assumptions and if these are consistent with the question that we ask, then that method has utility.



(With DTI) **(with CoD)**

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

As can be seen (Fig. 4), though CoD can detect linear lag influences, the strongly non-linear ones are missed. DTI detects these influences and exactly reproduces the synthetic network. Given the non-linear 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 putatively involved in these embryonic processes. Forty seven genes are expressed differentially between the ureteric bud and metanephric mesenchyme [47] and putatively involved in bud branching during kidney development. The expression data [8] temporally

profiles kidney development from day 10.5 dpc to the neonate stage. The influence network amongst these genes is shown below (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 below (Table. 2).

Some of the interactions that have been experimentally validated include the *Rara-Mapk1* [3], *Pax2-Gata3* [18] and *Agtr-Pax2* [53] 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.

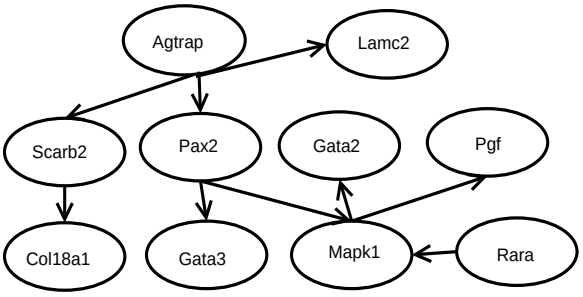


Fig. 5. Overall Influence network using DTI during early kidney development.

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 [41], in Fig. 6. This data represents the expression of various genes after T-cell activation using stimulation with phorbolester PMA and ionomycin. The dataset contains the profiles of about 58 genes over 10 time points with 44 replicate measurements for each time point.

Several of these interactions are confirmed in earlier studies [41, 16, 54, 43] and again point to the strength of DTI in recovering known interactions. The annotation of some of these genes are given in Table. 3. We note that the network of Fig. 6 shows the largest influence network (containing *Gata3*) that can be recovered. *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).

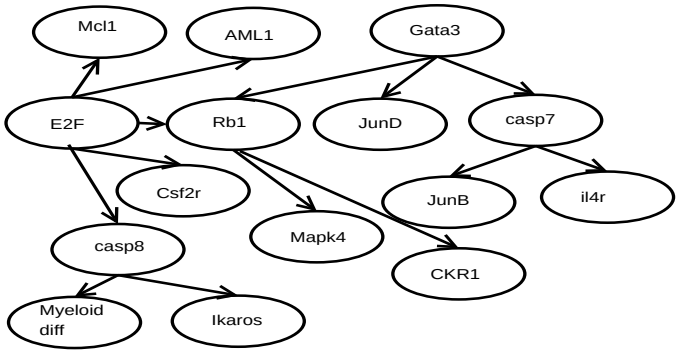


Fig. 6. DTI based T-cell network.

8.4. Phylogenetic conservation of TFBS effectors

A common approach to the determination of "functional" transcription factor binding sites in genomic regions is to look for motifs in conserved regions across various species. Here we focused on the interspecies conservation of TFBS (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 examine the *Gata3* gene promoter and find at least forty different transcription factors 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 [8, *http://spring.imb.uq.edu.au/*]. These "strength of influence" DTI values are first checked for signifi-

cance 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 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 as well as *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 transcription factors, 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, PPAR, SP1* have high influence via DTI and are expressed in embryonic kidney (Table. 4), apart from having conserved TFBS. 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 co-expression correlation with *Gata3* as well as conserved TFBS in the promoter region. The DTI criterion gave us no evidence of influence between these two TFs and *Gata3's* activity. This information coupled with the present evidence concerning the non-kidney specificity of *STE12* and *HP1*, presents an argument for the non-involvement 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, establish the DTI-based methodology as a common framework for both types of analysis.

8.5. Module TFs in co-regulated genes

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

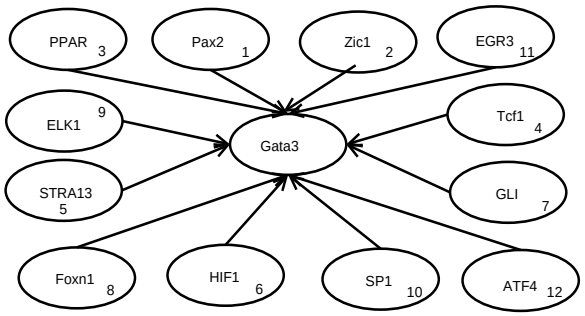


Fig. 7. Putative upstream TFs using DTI for the *Gata3* gene. The numbers in each TF oval represent the DTI rank of the respective TF.

Several tools (Genomatix [10], CREME [39], Toucan [1]) allow the inference of such transcription factor modules from sets of genes. However, the next logical question is if 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.

Table 2.
Functional annotations (*Entrez Gene*) of some of the genes co-expressed 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

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

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

Gene Symbol	Description	Expressed in Kidney
PPAR	peroxisome proliferator-activated receptor	Y
Pax2	Paired Homeobox-2	Y
HIF1	Hypoxia-inducible factor 1	Y
SP1	SP1 transcription factor	Y
GLI	GLI-Kruppel family member	Y
EGR3	early growth response 3	Y

To illustrate this application, genes that have expression in the developing Ureteric Bud (UB) in the kidney are examined. The inductive signals between the ureteric bud and metanephric mesenchyme causes 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 repository at: *http://www.informatics.jax.org/*), [48, 47] reveals some modules. The UB-specific genes as well as 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 $\sim$ 2000bp upstream to 200bp downstream from the transcription start site). The various promoters are then aligned and a search for significantly over-represented 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 [1] allows for the entire sequence of steps from sequence extraction to module searches. The list of all TFs in the various modules identified are listed in Table. 6.

Table 5. Genes expressed in the developing ureteric bud (day e10.5-11.0), as reported in 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 (Y/N) (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

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 has an important role in UB differentiation [18]. 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.

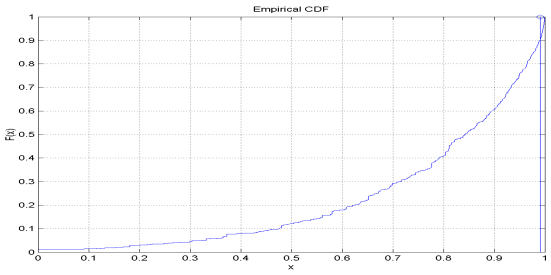


Fig. 8. Cumulative Distribution Function for bootstrapped $I(Pax2 \rightarrow Gata3)$. The true value of $I(Pax2 \rightarrow Gata3) = 0.9911$.

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, shown below, 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 [14, 18]. Thus, we can look at each module member for 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 (those that share some pattern of co-expression/co-regulation).

Extending this further, the strength and significance of the DTI can be found between every pair of TF and UB-specific gene 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 transcription factors in the modules and the co-expressed 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 transcription factors such as *SP1* have ubiquitous expression across most tissues [11, 44], and are not as informative as kidney-specific ones like *Pax2* [18] or *HNF4a* [49].

8.6. Higher-order MI and DI

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 amongst genes. Higher order MI [34, 35] has been used successfully for the discovery of interactions among triples of genes. Following work done in [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 hence DTI can be used to infer directional associ-

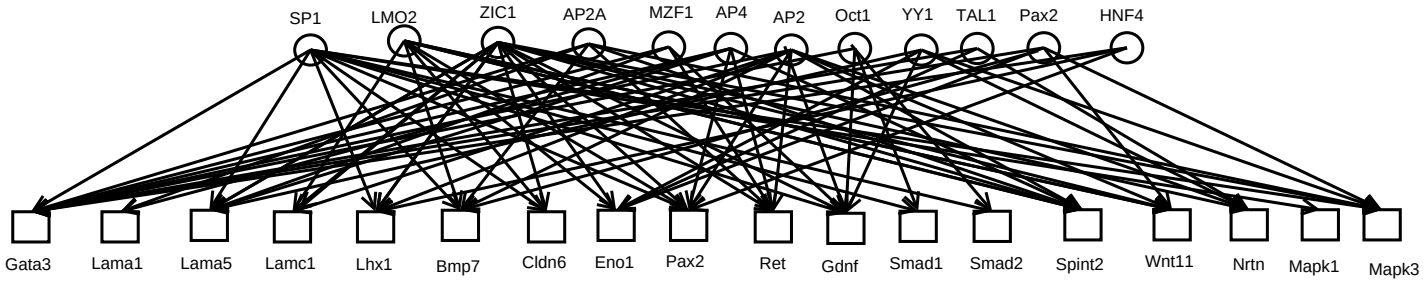


Fig. 9. A bipartite graph between the group of module TFs and genes co-expressed in the developing ureteric bud (MGI:e10.5-11.0).

ation 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 [7, 8, 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 the *Gdnf-Ret*, *Wnt*, and *Mapk* are implicated in ureteric bud differentiation [30, 13]. However, most studies have focussed on interaction within 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 co-operation/synergies between various pathways is 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.

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

			UB-Specificity & Citation (*http://symatlas.gnf.org/SymAtlas/*)
Gdnf	*Ret*	*Gata3*	Y [18]
Ret	*Bmp7*	*Gata3*	Y [13]
Pax2	*Gata3*	*Ret*	Y [9]
Ret	*Wnt11*	*Gdnf*	Y [30]
Pax2	*Wnt11*	*Gata3*	Y [18]
Pax2	*Ret*	*Gdnf*	Y [9,6]

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 non-linear interactions, we present two variants of DTI that can be used depending on context. One version, *unsupervised-DTI*, like traditional network inference, enables the discovery of influences (regulatory or non-regulatory) 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 as well as is general enough to accommodate arbitrary lag, non-linearity, and resolution of cycles, loops and direction.

We see that the above presented combination of supervised and unsupervised variants enable their

applicability to several important problems in bioinformatics (upstream TF discovery, module-gene interactions, and higher-order influence determination), some of which are presented in the results section. The network inference approach can also allow incorporation of additional biophysical knowledge - both pertaining to physical mechanisms as well as 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 and hence the DTI metric fits very naturally into such an integrative framework.

ACKNOWLEDGEMENTS

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 careful reading and offering several helpful insights to improve the quality of the manuscript.

References

1. 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.* 2003 Mar 15;31(6):1753-64.
2. Moonen C.T.W. (Ed), Bandettini P.A. (Ed), Functional MRI (Medical Radiology / Diagnostic Imaging), Springer, 2000.
3. Balmer JE, Blomhoff R., "Gene expression regulation by retinoic acid", *J. Lipid Res.* 2002 Nov;43:11:1773-808.
4. Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL.,"A Bayesian approach to reconstructing genetic regulatory networks with hidden factors", *Bioinformatics.* 2005 Feb 1;21(3):349-56.
5. Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *J. Roy. Statist. Soc. Ser. B.* 1995; 57:289-300.
6. 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.* 2001 Dec;128(23):4747-56.
7. 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.* 2004 Sep;15(9):2344-57.
8. 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.* 2005 Oct 17;23(2):159-71.
9. 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.* 2006 Dec 1;15(23):3420-8.
10. 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 U S A.* 2006 Apr 11;103(15):5682-7.
11. Cohen HT, Bossone SA, Zhu G, McDonald GA, Sukhatme VP.,"Sp1 is a critical regulator of the Wilms' tumor-1 gene", *J Biol Chem.* 1997 Jan 31;272(5):2901-13.
12. Cover T.M, Thomas J.A, "Elements of Information Theory", Wiley-Interscience, 1991.
13. Davies J, "Intracellular and extracellular regulation of ureteric bud morphogenesis", *Journal of Anatomy* 198 (3), 257264. (2001)
14. Dressler, G.R. and Douglas, E.C. , "Pax-2 is a DNA-binding protein expressed in embryonic kidney and Wilms tumor", *Proc. Natl. Acad. Sci. USA* 89: 1179-1183, 1992.
15. Effron B, Tibshirani R.J, An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability), Chapman & Hall/CRC, 1994.
16. Ezzat S, Mader R, Yu S, Ning T, Poussier P, Asa SL., "Ikaros integrates endocrine and immune system development", *J Clin Invest.* 2005 Apr;115(4):844-8.
17. Geweke J., "The Measurement of Linear Dependence and Feedback Between Multiple Time Series," *Journal of the American Statistical Association*, 1982, 77, 304-324. (With comments by E. Parzen, D. A. Pierce, W. Wei, and A. Zellner, and rejoinder)
18. 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.* 2006 Jan;133(1):53-61.
19. Gubner J. A., Probability and Random Processes for Electrical and Computer Engineers, Cambridge, 2006.
20. Hashimoto RF, Kim S, Shmulevich I, Zhang W, Bittner ML, Dougherty ER., "Growing genetic regulatory networks from seed genes",*Bioinformatics.* 2004 May 22;20(8):1241-7.
21. Hastie T, Tibshirani R, The Elements of Statistical

Learning, Springer 2002.
22. H. Joe., "Relative entropy measures of multivariate dependence", *J. Am. Statist. Assoc.*, 84:157164, 1989.
23. 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.* 2004 Dec;24(23):10263-76.
24. Kreiman G., "Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes"., *Nucleic Acids Res.* 2004 May 20;32(9):2889-900.
25. 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.* 1999 Feb;19(2):1558-68.
26. Miller E., "A new class of entropy estimators for multi-dimensional densities", *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.
27. Learned-Miller E., "Hyperspacings and the estimation of information theoretic quantities", UMass Amherst Technical Report 04-104, 2004.
28. Li H, Sun Y, Zhan M.,"Analysis of Gene Coexpression by B-Spline Based CoD Estimation",*EURASIP J Bioinform Syst Biol.* 2007;:49478.
29. MacIsaac KD, Fraenkel E., "Practical strategies for discovering regulatory DNA sequence motifs".,*PLoS Comput Biol.* 2006 Apr;2(4):e36.
30. 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.* 2003 Jul;130(14):3175-85.
31. 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.* 2006 Mar 20;7 Suppl 1:S7.
32. H. Marko, "The Bidirectional Communication Theory - A Generalization of Information Theory", *IEEE Transactions on Communications*, Vol. COM-21, pp. 1345-1351, 1973.
33. J. Massey, "Causality, feedback and directed information," in *Proc. 1990 Symp. Information Theory and Its Applications (ISITA-90)*, Waikiki, HI, Nov. 1990, pp. 303305.
34. Nemenman, F Shafee, and W Bialek. "Entropy and inference, revisited.", In TG Dietterich, S Becker, and Z Ghahramani, editors, *Advances in Neural Information Processing Systems* 14, Cambridge, MA, 2002. MIT Press.
35. Nemenman I,. "Information theory, multivariate dependence, and genetic network inference.", Technical Report NSF-KITP-04-54, KITP, UCSB, 2004.
36. Opgen-Rhein, R., and Strimmer K., "Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data", *Proc. of Fourth International Workshop on Computational Systems Biology, WCSB*, 2006.
37. I. Ovcharenko, M.A. Nobrega, G.G. Loots, and L. Stubbs, "ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes", *Nucleic Acids Research*, 32, W280-W286 (2004).
38. Paninski, L. , "Estimation of entropy and mutual information", *Neural Computation* 15: 1191-1254, 2003.
39. Papatsenko D., Levine M., "Computational identification of regulatory DNAs underlying animal development", *Nature Methods* 2, 529 - 534 (2005)
40. J. Ramsay, B. W. Silverman, Functional Data Analysis (Springer Series in Statistics), Springer 1997.
41. 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),1361-72, June 2004.
42. Rao A, Hero AO, States DJ, Engel JD, "Inference of biologically relevant Gene Influence Networks using the Directed Information Criterion", *Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2006.
43. 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.* 2004 Apr;24(7):2968-77.
44. Ryan G, Steele-Perkins V, Morris JF, Rauscher FJ, Dressler GR.,"Repression of Pax-2 by WT1 during normal kidney development", *Development.* 1995 Mar;121(3):867-75.
45. Schfer, J., and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks", *Bioinformatics*,Oct 2004.
46. Schneidman E, Still S, Berry MJ II, and Bialek W,"Network information and connected correlations", *Phys. Rev. Lett.*, 91, p. 238701, (2003)
47. Schwab K, Patterson LT, Aronow BJ, Luckas R, Liang HC, Potter SS., "A catalogue of gene expression in the developing kidney", *Kidney Int.* 2003 Nov; 64(5):1588-604.
48. Stuart RO, Bush KT, Nigam SK, "Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development", *Kidney International*,64(6),1997-2008,December 2003.
49. Taraviras S, Monaghan AP, Schtz G, Kelsey G.,"Characterization of the mouse HNF-4 gene and its expression during mouse embryogenesis", *Mech Dev.* 1994 Nov;48(2):67-79.
50. Venkataramanan R. , Pradhan S. S., "Source Coding With Feed-Forward: Rate-Distortion Theorems and Error Exponents for a General Source," *IEEE*

Transactions on Information Theory , vol.53, no.6, pp.2154-2179, Jun. 2007.
51. Willett R, Nowak R, "Complexity-Regularized Multiresolution Density Estimation", *Proc. Intl Symp. on Information Theory*, ISIT 2004.
52. Woolf PJ, Prudhomme W, Daheron L, Daley GQ, Lauffenburger DA., "Bayesian analysis of signaling networks governing embryonic stem cell fate decisions", *Bioinformatics*. 2005 Mar;21(6):741-53.
53. Zhang SL, Moini B, Ingelfinger JR., "Angiotensin II increases Pax-2 expression in fetal kidney cells via the AT2 receptor", *J Am Soc Nephrol*. 2004 Jun;15(6):1452-65.
54. Zhang, DH, Yang L, and Ray A., "Differential responsiveness of the IL-5 and IL-4 genes to transcription factor GATA-3", *J Immunol* 161: 3817-3821, 1998.

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, 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 $0 \leq I(X^N \rightarrow Y^N) \leq I(X^N;Y^N) < \infty$. 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 [19]: $\rho_{X^N;Y^N} = \Sigma_{X^N}^{-1/2}\Sigma_{X^NY^N}\Sigma_{Y^N}^{-1/2}$ 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 $H(X^N;Y^N) = -\frac{1}{2}\ln(2\pi e)^{2N}|\Sigma_{X^NY^N}|$, where $|A|$ is the determinant of matrix A, $\Sigma_{X^NY^N}$ denotes the covariance matrix, computed as $\Sigma_{X^NY^N} = \frac{1}{R-1}X^TY$, 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, $I(X;Y) = -\frac{1}{2}\ln(\frac{|\Sigma_{X^NY^N}|^2}{|\Sigma_{X^N}|.|\Sigma_{Y^N}|}) = -\frac{1}{2}\ln(1-\rho^2_{X^N;Y^N})$. Hence, a straightforward transformation is normalized MI, $\rho_{MI} = \sqrt{1-e^{-2I(X^N;Y^N)}} = \sqrt{1-e^{-2\sum_{i=1}^N I(X^N;Y_i|Y^{i-1})}}$. A connection with [22], 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 this approach, from [17]. Consider three random vectors $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, each of which are 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, $\delta_{YX|Z} = \sqrt{\frac{a_2^2}{a_1 a_3}}$ with, $a_1 = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$, $a_2 = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$, $a_3 = \Sigma_{XX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$.

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 mutual information to the directed information case, we have, $\rho_{DTI} = \sqrt{1-e^{-2\sum_{i=1}^N I(X^i;Y_i|Y^{i-1})}}$.

We recall the primary difference between MI and DTI, (note the superscript on X):
MI: $I(X^N;Y^N) = \sum_{i=1}^N I(X^N;Y_i|Y^{i-1})$.
DTI: $I(X^N \rightarrow Y^N) = \sum_{i=1}^N I(X^i;Y_i|Y^{i-1})$.

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 Section 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.