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. We demonstrate the conformity of the obtained interactions with literature as well as comparison with the coefficient of determination (CoD) method. Apart from network inference, the proposed framework enables the exploration of specific interactions, not just those revealed by data.

1. INTRODUCTION

Computational methods for inferring dependencies between genes [4,13,6] using probabilistic methods 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 techniques mostly yield networks of significant influences as 'observed/inferred' from the underlying structure of data. Alternatively, other biological data (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 potential limitations: biological significance and influence between *'any'* two variables of interest. Such an approach is increasingly necessary when we want 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 [5,26] can be interpreted as a directed version of mutual information, a criterion used quite frequently in other related work [13]. It turns out, as we will demonstrate, that the DTI gives a sense of directional association for the principled discovery of biological influence networks.

There are two main contributions of this work. Firstly, we present a short theoretical treatment of DTI and an approach to the supervised and unsupervised influence recovery problems, using microarray expression data. Secondly, we examine two sce-

*Corresponding author.

narios - 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.

2. 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, 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 [21]. Since transcription factors are also proteins (or their activated forms) which are in turn encoded for by other genes, we can consider the notion of an influence between a transcription factor gene and the target gene.

Below (Fig. 1) we give a characterization of what we mean by transcriptional regulatory networks. 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.

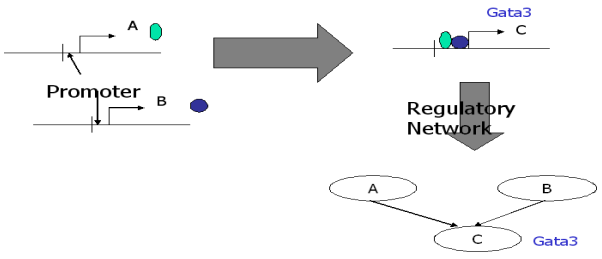


Fig. 1. A transcriptional regulatory network with genes A and B effect C. An example of C that we study here is the *Gata3* gene.

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 this paper we try to combine some of the other available biological data (protein-protein interaction data and phylogenetic conservation of binding sites across genomes) to build network topologies with a lower false positive rate of linkage.

3. PROBLEM SETUP

In this work, we also study the mechanism of gene regulation for genes, with the *Gata3* gene as an example. This gene has important roles in several processes in mammalian development [21], 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 [11, 12] would be 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). This work primarily addresses the following questions:

- 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 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.
- Biologists are also interested in network of relationships among genes expressed under a certain set of conditions, which uses several network inference procedures, such as Bayesian networks [4], MI [13] etc. However, there has been lack of a common framework to do both supervised *and* unsupervised *directed* network inference within these set-

tings to detect non-linear gene-gene interactions. We present Directed Information as a potential solution to 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).

3.1. Phylogenetic Conservation of Binding Sites

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 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. 2). Such information can be obtained from most genome browsers [20]. 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. 2). 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 encoding 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*. For illustration, we use the *Gata3* example in the rest of this paper.

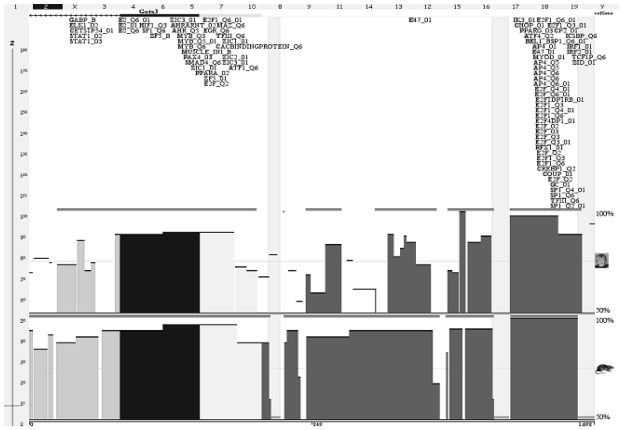


Fig. 2. TFBS conservation between Human, Mouse and Rat, upstream (x-axis) of *Gata3*, from *http://www.ecrbrowser.dcode.org*.

Another source of side information which becomes extremely useful in such scenarios is the biophysics of transcriptional regulation - this indicates that TFs binding at regulatory regions hardly do so alone but simultaneously participate in several interactions with proximal elements. Hence the presence of conserved TFs which are known binding partners (identified from protein interaction databases) increases the likelihood of functionality of that TF in transcriptional regulation. Our approach thus integrates several aspects:

- Identifying if any of the genes influence a target gene by coding for a transcription factor binding at the site discovered from conservation studies. This directed influence is captured using an influence metric (like directed information).
- Using phylogenetic information and protein-protein interaction to infer which binding sites upstream of a target gene may be functional.

4. 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 [13]. However, the absence of a 'causal' (or directed dependence) information theoretic metric has hindered the utilization of the full potential of information theory. In this work, we examine the applicability of such a metric - the Directed Information criterion (DTI) to the explicit inference of gene influence. This will enable us to potentially discover any directed non-linear relationship between genes of interest.

The DTI - which is a measure of the causal dependence between two N-length random processes $X \equiv X^N$ and $Y \equiv Y^N$ is given by [22]:

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

Here, Y^n denotes $(Y_1, Y_2, .., Y_n)$, i.e. a segment of the realization of a random sequence Y and $I(X^N; Y^N)$ is the Shannon mutual information [28].

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}). We do this for every instant $n \in (1, 2, \ldots, N)$ in the N - length expression time series. Thus, we find the influence relationship between genes X and Y for every instant during the evolution of their individual 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 Shannon entropy of X and the conditional entropy of X given Y, respectively. Using this definition of mutual information, the Directed Information can be expressed in terms of individual and joint entropies of X and Y. One way to estimate entropy is to use marginal and joint histograms, but there are problems both due to computational complexity as well as with moderate sample size. Especially in a microarray expression setting (where we have only a modest number of sample points per gene), it would be useful to examine an alternative strategy for entropy estimation which uses a data-dependent binning approach. One such method to find the entropy of the random variables X^N and Y^N uses the Darbellay-Vajda algorithm [7]. In this approach, an adaptive partitioning of the observation space is used to estimate the probability densities as well as the entropies of the random variables.

Briefly, the Darbellay-Vajda procedure for entropy estimation proceeds as follows (more details can be found in [23]):

$$\begin{aligned} I(X^N \to Y^N) &= \sum_{n=1}^{N} [H(X^n|Y^{n-1}) - H(X^n|Y^n)] \\ &= \sum_{n=1}^{N} [I(X^n; Y^n) - I(X^n; 0Y^{n-1})] \qquad (2) \end{aligned}$$

- For evaluating the DTI in expression (2), we need to evaluate the expressions $I(X^n; Y^n)$ and $I(X^n; 0Y^{n-1})$ in each term of the sum. For the evaluation of $I(X^n; Y^n)$, we thus have an n-dimensional list for X^n and Y^n respectively.
- Transform the vectors $X^n \equiv (X_1, X_2, \ldots, X_n)$, $Y^n \equiv (Y_1, Y_2, \ldots, Y_n)$ to $(U_i, V_i) \equiv (j : X_{(j)} = X_i, k : Y_{(k)} = Y_i), \forall 1 \leq i \leq n$ where $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$, $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)})$ are the rank-ordered versions of $(X_1, X_2, \ldots, X_n)$, $(Y_1, Y_2, \ldots, Y_n)$. Thus the sample (observation) space $((U, V))$ is a $2D$ representation of the ranks of X^n and Y^n. This is an ordinal sampling step. We note that $I(U, V) = I(X^n, Y^n)$.
- In the $U-V$ co-ordinate plane, a dyadic partitioning of the sample space is iteratively done until the sample distribution of each cell is not significantly different than random (i.e. conditionally independent). Once the sample distribution in a cell achieves independence, it (the cell) is not split any further.
- Hence, if there are K partitions in the observation space, and the k^{th} cell has n_k samples, the mutual information is estimated as $I_{U,V} = I(X^n, Y^n) = \Sigma_{k=1}^{K} \frac{n_k}{n} \times$

$log\big(\frac{n_k/n}{(L_k^U/n)(L_k^V/n)}\big)$. We note that $\frac{L_k^U.L_k^V}{n^2}$ is the $2D$-hypervolume of the k^{th} cell.

- We note that the presence of biological/technical replicates (as is available in microarray data) would create many more samples from which to obtain entropy estimates.

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 information estimates ($I(X^n; 0Y^{n-1})$, and $I(X^n; Y^n)$) computed above into the above expression (2). However, it is preferred to have a normalized version of this metric (lying between $[0, 1]$) for a comparison of the strengths of relationships between other genes. Also, it is essential to consider a notion of significance of the obtained DTI measure. We thus perform bootstrapping of every estimate of the DTI and if the value of DTI is significant (p value = 0.05), we accept the notion of an influence between genes X and Y. Below (Sec: 7), we have indicated the sequence of steps to estimate the significance of an influence between *Pax2* and *Gata3*.

The steps for normalizing the DTI measure as well as estimating significance with respect to a null DTI distribution are given in the following sections.

5. A NORMALIZED DTI MEASURE

In this section, we derive an expression for a 'normalized DTI coefficient'. This is useful for a meaningful comparison across different criteria during network inference. 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 [24]: $\rho_{X^N;Y^N} = \Sigma_{X^N}^{-1/2} \Sigma_{X^N;Y^N} \Sigma_{Y^N}^{-1/2}$ 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 $|A|$ is the determinant of matrix A, Σ denotes the covariance matrix.

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, $I(X; Y) = -\frac{1}{2} \ln(\frac{|\Sigma_{X^N Y^N}|^2}{|\Sigma_{X^N}|.|\Sigma_{Y^N}|}) = -\frac{1}{2} \ln(1 - \rho^2_{X^N;Y^N})$. Hence, a straightforward transformation is normalized MI, $\rho_{MI} = \sqrt{1 - e^{-2I(X;Y)}} = \sqrt{1 - e^{-2\sum_{i=1}^{N} I(X^N; Y_i | Y^{i-1})}}$. A connection with [15], 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 [9]. 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 the next two sections.

6. KERNEL DENSITY ESTIMATION (KDE)

The goal in density estimation is to find a probability density function $\hat{f}(z)$ that approximates the underlying density $f(z)$ of the random variable Z. Under certain regularity conditions, the kernel density estimator $\hat{f}_h(Z)$ at the point z is given by $\hat{f}_h(Z) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{z_i - z}{h})$, with n being the number of samples $z_1, z_2, \ldots, z_n$ from which the density

is to be estimated, h is the bandwidth of a kernel $K(\bullet)$ that is used during density estimation.

A kernel density estimator at z works by weighting the samples (in $(z_1, z_2, \ldots, z_n)$) around z by a kernel function (window) and counts the relative frequency of the weighted samples within the window width. As is clear from such a framework, the choice of kernel function $K(\bullet)$ and the bandwidth h determines the fit of the density estimate.

Some figures of merit to evaluate various kernels are the asymptotic mean integrated squared error (AMISE), bias-variance characteristics and region of support [8]. It is preferred that a kernel have a finite range of support, low AMISE and a favorable bias-variance tradeoff. The bias is reduced if the kernel bandwidth (region of support) is small, but has higher variance because of a small sample size. For a larger bandwidth, this is reversed (ie large bias and smaller variance). Under these requirements, the Epanechnikov kernel has the most of these desirable characteristics - i.e. a compact region of support, the lowest AMISE compared to other kernels, and a favorable bias variance tradeoff [8].

The Epanechnikov kernel is given by:

$$K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1).$$

with $I(\bullet)$ being the indicator function conveying a window of width spanning $[-1, 1]$ centered at 0. An optimal choice of the bandwidth is $h = 1.06 \times \hat{\sigma}_z \times n^{-1/5};$, following [14]. Here $\hat{\sigma}_z$ is the standard error from the bootstrap DTI samples $(z_1, z_2, \ldots, z_n)$.

Hence the kernel density estimate for the bootstrapped DTI (with $n = 1000$ samples), $Z \triangleq \hat{I}_B(X^N \rightarrow Y^N)$ becomes,
$\hat{f}_h(Z) = \frac{1}{nh}\sum_{i=1}^{n}\frac{3}{4}[1 - (\frac{z_i - z}{h})^2]I(\left|\frac{z_i - z}{h}\right| \leq 1)$ with $h \approx 2.67\hat{\sigma}_z$ and $n = 1000$. We note that $\hat{I}_B(X^N \rightarrow Y^N)$ is obtained by finding the DTI for each random permutation of X, Y time series, performing this permutation B times, and obtaining a estimate of the density over these B permutations.

7. BOOTSTRAPPED CONFIDENCE INTERVALS

Since we do not know the true distribution of the DTI estimate, we find an approximate confidence interval for the DTI estimate $(\hat{I}(X^N \rightarrow Y^N))$, using bootstrap [19]. We denote the cumulative distribution function (over the Bootstrap samples) of $\hat{I}(X^N \rightarrow Y^N)$ by $F_{\hat{I_B}(X^N \rightarrow Y^N)}(\hat{I_B}(X^N \rightarrow Y^N))$, Figure 3. Let the mean of the bootstrapped null distribution be $I_B^*(X^N \rightarrow Y^N)$. We denote by $t_{1-\alpha}$, the $(1-\alpha)^{th}$ quantile of this distribution i.e. $\{t_{1-\alpha} : P([\frac{\hat{I_B}(X^N \rightarrow Y^N) - I_B^*(X^N \rightarrow Y^N)}{\hat{\sigma}}] \leq t_{1-\alpha}) = 1 - \alpha\}$. Since we need the real $\hat{I}(X^N \rightarrow Y^N)$ to be significant and close to 1, we need $\hat{I}(X^N \rightarrow Y^N) \geq [I_B^*(X^N \rightarrow Y^N) + t_{1-\alpha} \times \hat{\sigma}]$, with $\hat{\sigma}$ being the standard error of the bootstrapped distribution, $\hat{\sigma} = \sqrt{\frac{[\Sigma_{b=1}^{B}\hat{I}_b(X^N \rightarrow Y^N) - I_B^*(X^N \rightarrow Y^N)]^2}{B-1}}$; B is the number of Bootstrap samples.

For the *Pax2-Gata3* interaction, we show the kernel density estimate of the bootstrapped histogram using the Epanechnikov kernel (Fig. 3) as well as the position of the true DTI estimate in relation to the overall histogram. With the obtained kernel 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.

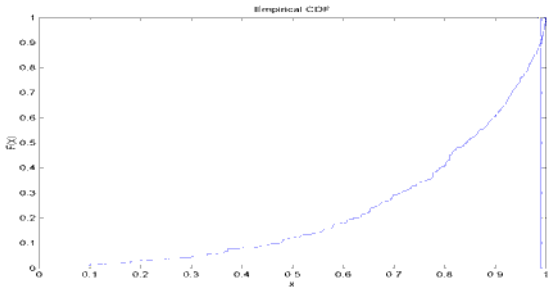


Fig. 3. Cumulative Distribution Function for bootstrapped $I(Pax2 \rightarrow Gata3)$. The true $I(Pax2 \rightarrow Gata3) = 0.9911$.

8. 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 for (*supervised-DTI*) is as follows:

- Identify the G key genes based on required phenotypical characteristic using fold change studies. Preprocess the gene expression profiles by normalization and cubic spline interpolation. We now assume that there are N

points for each gene. Bin each of the expression profiles into K quantiles (here, we use $K = 4$), thus building a joint histogram. The granularity of sampling can be an issue during entropy estimation, hence the Darbellay-Vajda method can also be used here. We note that the presence of probe-level or sample replicates greatly enhance the accuracy of the entropy estimation step.

- For each pair of genes A_i and B among these G genes :
  {
  - Look for a phylogenetically conserved binding site of protein 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, $DTI(A_i, B) = \sqrt{1 - e^{-2I(A_i^N \rightarrow B^N)}}$.
  - Bootstrapping over several permutations of the data points of A_i and B yields a null distribution (using KDE) 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 'affector'. This search is done for every gene A_i among these G genes $((A_1, A_2, \ldots, A_G))$.

  }
- We observe that both phylogenetic information is inherently built into the influence network inference step above.

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 the false discovery rate. We note that it is fairly simple to include some apriori biological knowledge (if a subset of upstream TFs at the promoter are already known, either experimentally or from other sources) - a search among the binding partners of these known TFs can reduce the set of potential effectors and reduce the complexity of the unsupervised procedure. Another element that has been added is the control of false discovery rate (FDR) [27] to screen each of the $G(G-1)$ hypotheses (both directions) during network discovery amongst G genes.

Table 1. Comparison of various network inference methods.

Method	Resolve Cycles	Non-linear framework	Search for interaction	Non-parametric framework
SSM [1]	Y	N	N	N
CoD [3]	N	N	Y	N
GGM [6]	N	Y	N	N
DTI [5]	Y	Y	Y	Y

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 the only metric which can recover interactions under all these considerations.

9. 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 three 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 binding motif sequences conserved 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.

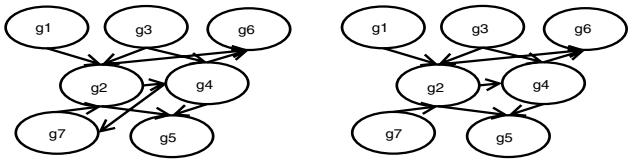
Before proceeding, we examine the performance of this approach on synthetic data.

9.1. Synthetic Network

A synthetic network is constructed in the following fashion: We assume that there are two genes g_1 and g_3 which drive the remaining genes of a seven gene network. The evolution equations are as below:

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

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 [3] 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 non-linear ones are missed. DTI detects these influences and almost 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.

9.2. 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 [25] and putatively involved in bud branching during kidney development. The expression data [10] temporally profiles kidney development from day 10.5 dpc to the neonate stage. The influence amongst these genes is

shown below (Fig. 5). Several of the presented interactions are biologically validated but 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).

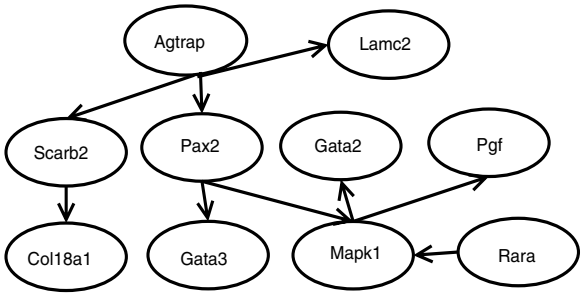


Fig. 5. Overall Influence network using DTI during early kidney development.

Some of the interactions that have been experimentally validated include the *Rara-Mapk1* [18], *Pax2-Gata3* [16] and *Agtr-Pax2* [17] 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 apriori. Such a scenario will be explored later (Sec: 9.4). 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.

An important shortcoming of most gene network inference approaches is that these relationships are detected based on mRNA expression levels alone. To understand these interactions with greater fidelity, there is a need to integrate other data sources corresponding to phosphorylation, dephosphorylation as well as other post-transcriptional/translational activities, including miRNA activity.

9.3. 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 [1], in Fig. 6. This data looks at the expression of various genes after T-cell activation using stimulation with phorbolester PMA and ionomycin. This data has the profiles of about 58 genes over 10 time points with 44 (34+10) replicate measurements for each time point.

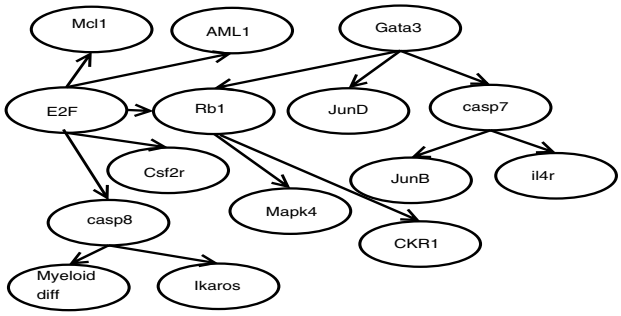


Fig. 6. DTI based T-cell network.

Several of these interactions are confirmed in earlier studies [1, 29, 30, 31] 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).

9.4. 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 inter-species conservation of TFBS (Fig. 2) 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. 2, we examine the *Gata3* gene promoter and find atleast 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.

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

Using *supervised DTI*, we examined the strength of influence from each of the TF-encoding genes (A_i) to *Gata3*, based on expression level [10, *http://spring.imb.uq.edu.au/*]. These "strength of influence" DTI values are 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 highest influence on *Gata3* expression (Fig. 7). Obviously, this information is far from complete, because of examination only at the mRNA level for both effector as well as *Gata3*.

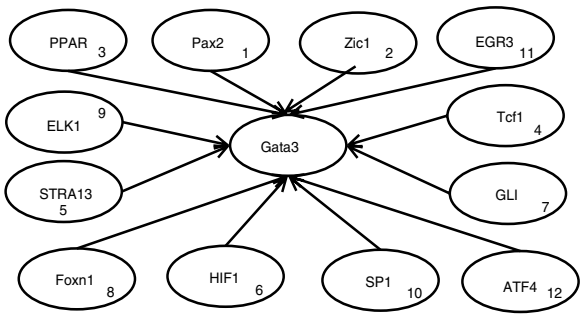


Fig. 7. Putative upstream TFs using DTI for the *Gata3* gene. The numbers in each TF oval represent the DTI rank of the respective TF.

Table 4 shows the embryonic kidney-specific expression of the TFs from 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.

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

As an additional step, we also 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 to TFs and *Gata3's* activity. We believe that this information coupled with the present evidence concerning the non-kidney specificity of *STE12* and *HP1*, present some argument for the non-involvement of these TFs in kidney specific regulation of *Gata3*. Hopefully, these findings would guide a more focused experiment to identify the key TFs involved in *Gata3* activity.

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 the same framework for both these purposes as well as is general enough to accommodate arbitrary lag, non-linearity, 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), some of which are presented in the Results section. The network inference approach can also alow 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 in 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 also grateful to the reviewers for having helped us to improve the quality of the manuscript.

References

1. Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, Gaiba A, Wild DL, Falciani F, 11Modeling T-cell activation using gene expression profiling and state-space models", *Bioinformatics*, 20(9),1361-72, June 2004.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. on Information Theory*, vol. 45, pp. 1315–1321, May 1999.
8. Hastie T, Tibshirani R, The Elements of Statistical Learning, Springer 2002.
9. 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)
10. 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.
11. Kreiman G., "Identification of sparsely distributed clusters of cis-regulatory elements in sets of coexpressed genes".,*Nucleic Acids Res*. 2004 May 20;32(9):2889-900.
12. MacIsaac KD, Fraenkel E., "Practical strategies for discovering regulatory DNA sequence motifs".,*PLoS Comput Biol*. 2006 Apr;2(4):e36.

13. 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.
14. J. Ramsay, B. W. Silverman, Functional Data Analysis (Springer Series in Statistics), Springer 1997.
15. H. Joe., "Relative entropy measures of multivariate dependence"., *J. Am. Statist. Assoc.*, 84:157164, 1989.
16. 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.
17. 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.
18. Balmer JE, Blomhoff R., "Gene expression regulation by retinoic acid".,*J. Lipid Res.* 2002 Nov;43:11:1773-808.
19. Effron B, Tibshirani R.J, An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability), Chapman & Hall/CRC, 1994.
20. 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).
21. 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.
22. J. Massey, "Causality, feedback and directed information," in *Proc. 1990 Symp. Information Theory and Its Applications (ISITA-90)*, Waikiki, HI, Nov. 1990, pp. 303305.
23. Hudson, J.E., "Signal Processing Using Mutual Information", *Signal Processing Magazine*,23(6):50-54, Nov. 2006.
24. Gubner J. A., Probability and Random Processes for Electrical and Computer Engineers, Cambridge, 2006.
25. 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.
26. H. Marko, "The Bidirectional Communication Theory - A Generalization of Information Theory", *IEEE Transactions on Communications*, Vol. COM-21, pp. 1345-1351, 1973.
27. 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.
28. Cover T.M, Thomas J.A, "Elements of Information Theory", Wiley-Interscience, 1991.
29. 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.
30. 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.
31. 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.