

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.

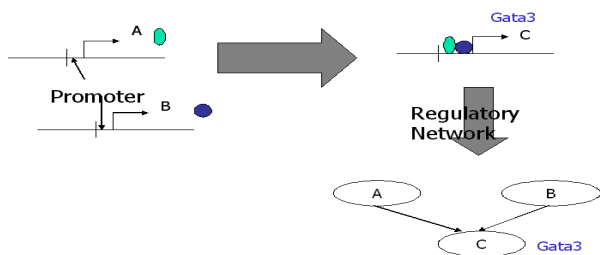


Fig. 1. A transcriptional regulatory network with genes A and B effect C. An example of C that we study here is the *Gata3* gene.

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. Un-supervised 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.

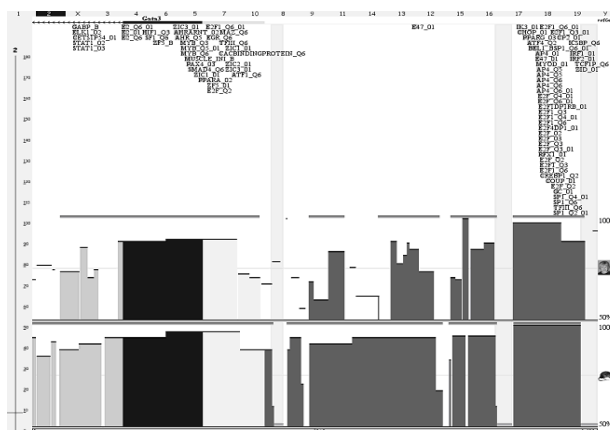


Fig. 2. TFBS conservation between Human, Mouse and Rat, upstream (x-axis) of *Gata3*, from <http://www.ecrbrowser.dcode.org>.

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]:

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

Here, Y^n denotes (Y_1, Y_2, \dots, 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 $n \in (1, 2, \dots, N)$ 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 microar-

ray 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]):

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

- 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 $X^n \equiv (X_1, X_2, \dots, X_n)$, $Y^n \equiv (Y_1, Y_2, \dots, Y_n)$ to $(U_i, V_i) \equiv (j : X_{(j)} = X_i, k : Y_{(k)} = Y_i), \forall 1 \leq i \leq n$ where $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$, $(Y_{(1)}, Y_{(2)}, \dots, Y_{(n)})$ are the rank-ordered versions of (X_1, X_2, \dots, X_n) , (Y_1, Y_2, \dots, 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 $I_{U,V} = I(X^n, Y^n) = \sum_{k=1}^K \frac{n_k}{n} \times$

