INFERENCE OF BIOLOGICALLY RELEVANT GENE INFLUENCE NETWORKS USING THE DIRECTED INFORMATION CRITERION

Arvind Rao [2,3,4], *Alfred O. Hero* [1,2,4,6], *David J. States* [2,5], *James Douglas Engel* [3]

Departments of [1]Biomedical Engineering, [2]Bioinformatics, [3]Cell and Developmental Biology, [4]Electrical Engineering and Computer Science,[5]Human Genetics,[6]Statistics
The University of Michigan, Ann Arbor, MI

ABSTRACT

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 us to use 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 probabilistic network inference methodology, based on the Directed information criterion, which incorporates the biology of transcription within the framework, so as to enable experimentally verifiable inference. We use a publicly available embryonic kidney microarray dataset to demonstrate our results on the regulation of the *Gata2/Gata3* genes.

Keywords: Transcriptional regulation, phylogeny, directed information, protein-protein interaction, Transcription factor Binding sites (TFBS).

1. INTRODUCTION

Computational methods for exploiting probabilistic dependencies between gene expression, proteomic [8] have been exploited for quite some time now. However their biological significance 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. What if we were interested in the influences on a certain variable 'A' but our prospective network inference technique was unable to recover them? In this work we propose a similar probabilistic technique with an eye on two of these potential limitations: biological significance and influence between 'any' variables of choice/interest.

The probabilistic method that we propose builds on an information theoretic criterion referred to as the Directed Information (DTI). The DTI [5] can be loosely interpreted as a directed version of mutual information, a criterion used quite frequently in other work [9]. It turns out, as we will demonstrate, that the DTI gives a sense of both directionality as well as dependence for the inference of influence networks.
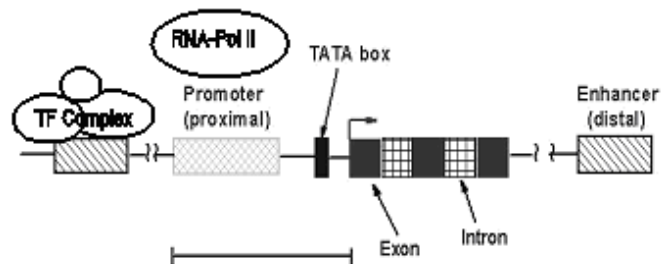

Fig. 1. Schematic of Transcriptional Regulation.

2. GENE NETWORKS

Below we give a characterization of what we mean by transcriptional regulatory networks [4]. As the name suggests, gene A is connected by a link to gene B if a product of gene A, say protein A, is involved in the transcriptional regulation of gene B. This might mean that protein A is involved in the formation of the complex which binds at the basal transcriptional machinery of gene B to drive gene B regulation. This is indicated below:

As can be seen, the components of the Transcription Factor (TF) Complex, shown in Fig. 1, are the products of several genes. Therefore, the incorrect inference of a transcriptional regulatory network can led to several 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. To make the inference of these networks relevant for a biologist to design useful experiments it would seem imperative that we incorporate biological knowledge to an extent suitable for making such network inference meaningful. In this paper we try to combine some of the other available 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.
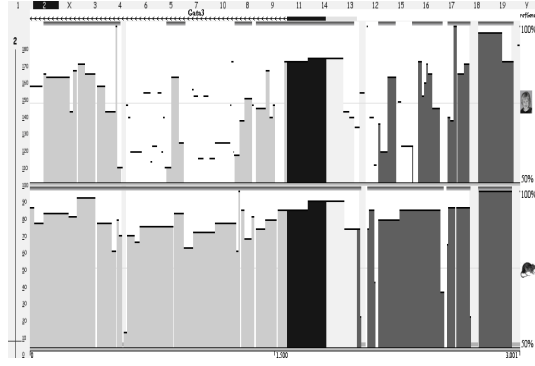
Fig. 2. TFBS conservation between Human and Rat, upstream (x-axis) of *Gata3*.

3. FINDING PHYLOGENETIC CONSERVATION OF BINDING SITES

As already 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 motifs might have appeared over the evolutionary time period due to insertions, mutations, deletions etc in the 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 occurred. This however, does not suggest that those other binding sites (TFBS) have no functional role. Since we are interested in the mechanism of regulation of the *Gata2/Gata3* genes (which are known to be implicated in mammalian nephrogenesis), we examine their promoter regions for phylogenetically conserved TFBS (Fig. 2). Such information can be obtained from most genome browsers [2]. 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). It is extremely important to select only a subset of the TFs that could bind at these sites for experimental testing, 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 Directed information. 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. The data source for directed information inference is an independent data source - a public repository of kidney microarray data (http://genet.chmcc.org). From here onwards, for the purpose of illustration, we continue with the *Gata3* example to demonstrate our results.

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 potential functionality of that TFBS for transcriptional regulation. Our approach thus involves two distinct components:

- Using phylogenetic information and protein-protein interaction to infer which binding sites upstream of a target gene may be functional.
- Identifying if any of the genes (identified via fold change analysis of microarray data), influence a target gene by coding for a transcription factor binding at the site discovered in step 1. This causal influence is captured using the Directed information criterion.

4. USING DIRECTED INFORMATION FOR INFERENCE OF NETWORK TOPOLOGY

Traditionally, there has been a lot of work exploring the feasibility of using Mutual information as a method to infer the conditional dependence/influence among genes by exploring the structure of the joint distribution of the gene expression profiles [9]. However, the absence of a 'causal' information theoretic metric has prevented us from exploitation 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 uncover any meaningful relationship between the DTI metric to the known causal influences among genes.

The DTI (for a lag of 1) - which is a measure of the causal dependence between two random processes X and Y is given by [6]:

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

Here, Y^n denotes $(Y_1, Y_2, .., Y_n)$, i.e. a segment of the realization of a random sequence Y and I(X;Y) is the Shannon mutual information . As already known, $I(X;Y) = H(X) - H(X|Y)$, with $H(X)$ and $H(X|Y)$ being the Shannon entropy of X and the conditional entropy of X given Y, respectively. Using this definition of mutual information, the Directed Information simplifies to,

$$I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} [H(X^n | Y^{n-1}) - H(Y_n | Y^{n-1})]$$

$$= \sum_{n=1}^{N} \{[H(X^n, Y^{n-1}) - H(Y^{n-1})] - [H(Y_n, Y^{n-1}) - H(Y^{n-1})]\}$$

$$(2)$$

Using (2), the Directed information is expressed in terms of individual and joint entropies of X and Y (these can be estimated using standard entropy estimation methods).

The entropy of a gene profile $X = (X_1, X_2, \ldots, X_n)$ is estimated by the following procedure:

- Each gene profile is normalized to have mean 0 and unit variance. Every normalized gene profile is then quantized into K quantiles (bins) with the control points $c = [c_0, c_1, \ldots, c_K]$, with c_i denoting the ith quantile.

- For any two given genes, we estimate the bivariate histogram by:

$$p_{XY}(i,j) = \frac{1}{n} \sum_{i,j} Ind(c_{i-1} < x \leq c_i,$$

$$c_{j-1} < y \leq c_j),$$

Here, the indicator function $Ind(c_{i-1} < X \leq c_i)$ is defined as: $Ind(c_{i-1} < X \leq c_i) = 1$ if X lies in the ith bin, with 0 otherwise. Note: $i \in (1, \ldots, K)$

- The individual entropy is computed by estimation of a univariate histogram:

$$p_X(i) = \frac{1}{n} \sum_i Ind(c_{i-1} < x < c_i)$$

- The various entropies are computed using:

$$H(X) = -\sum_i p_X(x) \ln p_X(x)$$

$$H(X,Y) = -\sum_{i,j} p_{XY}(x,y) \ln p_{XY}(x,y)$$

$$H(X|Y) = H(X,Y) - H(Y)$$

We then 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 influence between genes A and B. Below, we have indicated one such DTI distribution generated by bootstrapping, to estimate the significance of an influence between *PPAR-RXR* and *Gata3*.

Thus, our proposed approach is as follows:

- Identify the G key genes based on required phenotypical characteristic using fold change studies [7]. 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. We note that the presence of probe-level or sample replicates greatly enhance the accuracy of the entropy estimation step.
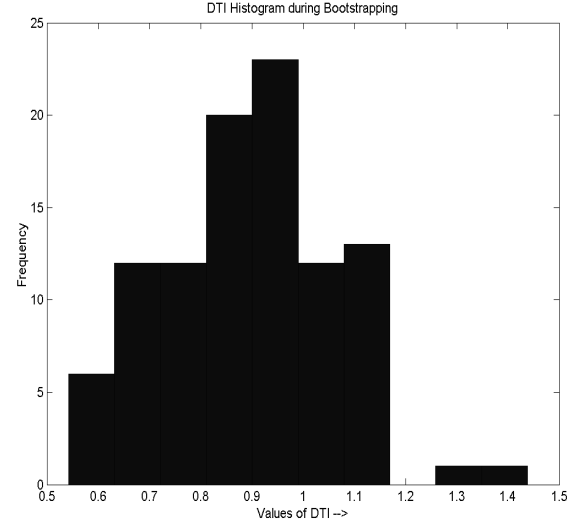
Fig. 3. DTI histogram from Bootstrapping, for inferring influence between *PPAR-RXR* and *Gata3*.

- For each pair of genes A and B among these G genes :
  {

  – Look for a phylogenetically conserved binding site of protein encoded by gene A in the upstream region of gene B.

  – Find $DTI(A,B) = I(A^N \rightarrow B^N)$, using equation (1).

  – If Bootstrapping over several permutations of the data points of A and B yields that $DTI(A,B)$ is within the 99% confidence interval (CI) from the histogram obtained from bootstrapping, infer a potential influence from A to B.

  – Every gene A which is potentially influencing B is an 'affector'.

  }

- If the product of A has binding partners among the other affectors identified till the present, use the 95% bootstrapped CI - this reduces the stringency for gene A, since we know that we have further evidence for the role of A in B's regulation.

- The search over all pairs of genes among these G genes yields an influence network. We observe that both phylogenetic information as well as the biophysics during regulation is inherently built into the influence network inference step above.
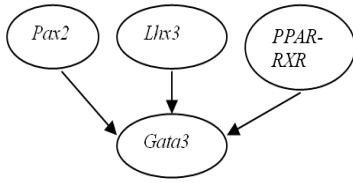
Fig. 4. Influence network using DTI for the *Gata3* gene

5. RESULTS

While examining the upstream region of *Gata3*, we found binding sites [2] for the TFs *Lhx3, Pax2, PPAR-RXR*, among others. These genes are also seen to be over-expressed in microarray experiments of the embryonic kidney. *Gata3* is involved in early kidney formation hence the presence of these TFBS in the promoter element of *Gata3* is of interest. Therefore, we ask if there is additional confirmation for any of these three genes affecting our target gene (*Gata3*), as obtained from DTI on available microarray data (Fig. 4).

We observe that the genes coding for these particular TFs indeed have an influence on *Gata3* expression. This helps us hypothesize that these three TFs are possibly functional, since they have an influence, are present in the *Gata3* promoter, and are expressed in early embryonic kidney. As observed from Fig.2 we can now concentrate on these TFs instead of the large number of TFBS observed only from phylogenetic analysis.

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, this would guide a more focused experiment to identify the key TFs involved in *Gata3* activity.

Our observation has been that other methods for network inference, such as in [1,3], rely on the probabilistic dependencies in the acquired data, and can be curated for biological significance only after inference is done. Though the influences recovered from these approaches may be biologically relevant, we are not aware of any previous studies to actively infer the nature of influence between two given genes.

6. CONCLUSIONS

We have presented the notion of Directed information as a reliable criterion for the inference of influence in gene networks. The procedure for inference incorporates biological knowledge in the form of Transcription factor binding site conservation as well as biophysics of the transcriptional regulation. We find that instead of only 'recovering' influences from the structure of high throughput data, one can actively look for the strength of influence via directed information. 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.

8. REFERENCES

[1] Schafer, J., and K. Strimmer. "An empirical Bayes approach to inferring large-scale gene association networks". Bioinformatics 21: 754-764, 2005.

[2] G.G. Loots and I. Ovcharenko, "rVISTA 2.0: evolutionary analysis of transcription factor binding sites", Nucleic Acids Research, 32(Web Server Issue), W217-W221 (2004)

[3] A Rao, AO Hero, DJ States, JD Engel, "Inferring Time Varying Network Topologies from Gene Expression data", IEEE Genomic Signal Processing and Statistics (GENSIPS) 2005.

[4] Alberts, Bruce; Johnson, Alexander; Lewis, Julian; Raff, Martin; Roberts, Keith; Walter, Peter ,"Molecular Biology of the Cell", New York: Garland Publishing; 2002.

[5] Ramji Venkataramanan and S. Pradhan, "Directed Information for Communication Problems with Common Side Information and Delayed Feedback/Feedforward", Proc. of the 43rd Annual Allerton Conference (Monticello, IL), Oct. 2005.

[6] J. Massey, Causality, "Feedback and Directed Information", Proceedings of the 1990 Symposium on Information Theory and its Applications (ISITA-90), pp. 303-305, 1990.

[7] Stuart RO, Bush KT, Nigam SK, "Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development", Kidney Int. 2003 Dec;64(6):1997-2008.

[8] Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP.,"Causal protein-signaling networks derived from multiparameter single-cell data". Science. 2005 Apr 22;308(5721):523-9.

[9] Hartemink AJ.,"Reverse engineering gene regulatory networks", Nature Biotechnology 23, 554 - 555 (2005).