MANIFOLD EMBEDDING FOR UNDERSTANDING MECHANISMS OF TRANSCRIPTIONAL REGULATION

Arvind Rao [b,c,d], *Alfred O. Hero* [a,b,d,f], *David J. States* [b,e], *James Douglas Engel* [c]

Departments of [a]Biomedical Engineering, [b]Bioinformatics, [c]Cell and Developmental Biology,
[d]Electrical Engineering and Computer Science,[e]Human Genetics,[f]Statistics,
The University of Michigan, Ann Arbor, MI

ABSTRACT

In this paper we propose a manifold embedding methodology to integrate heterogeneous sources of genomic data for the purpose of interpretation of transcriptional regulatory phenomena and subsequent visualization. Using the *Gata3* gene as an example, we ask if it is possible to determine which genes (or their products) might be potentially involved in its tissue-specific regulation - based on evidence obtained from various available data sources. Our approach is based on co-embedding of genes onto a manifold wherein the proximity of neighbors is influenced by the probability of their interaction as reported from diverse data sources - i.e. the stronger the evidence for that gene-gene interaction, the closer they are.

1. INTRODUCTION

Computational inference of transcriptional regulatory networks from diverse data has proved to be a bigger challenge than previously imagined. The gold standards for each data source are highly variable, and considering the diversity of interactions that each experimental or computational method aims to recover, their meaningful integration for the purpose of understanding underlying phenomena is a non-trivial task. For this study, we examine three kinds of data sources, two of which are experimentally derived (protein-protein interaction assays, phylogenetic conservation of Transcription Factor Binding sites (TFBS)) and the third is a computational measure (Directed Information) [3] for inferring interactions. Our objective is to demonstrate that not only is this method scalable to as many kinds of 'relevant' data sources but also encompass both experimental and computational measures of association. Our approach is to construct an interaction probability matrix between K genes under consideration. This matrix is a $K \times K$ symmetric matrix with $P(i,j) = P(j,i) = P(Z_{i,j} = 1)$, the probability that there is a 'true' functional interaction between the genes i and j, denoted by the event $Z_{i,j} = 1$. This true interaction depends on the probability that the l^{th} data source confirms this interaction (i.e. $Z^l_{i,j} = 1$). If we have L (=3, here) different data sources, we can write this as:

$$p_{i,j} = P(Z_{i,j} = 1 | Z^1_{i,j} = 1, Z^2_{i,j} = 1, \ldots, Z^L_{i,j} = 1) \propto \prod_{l=1}^{L} P(Z_{i,j} = 1 | Z^l_{i,j} = 1) \qquad (1)$$

Thus, the existence of a 'true' functional relation between two genes i and j, depends on $p_l = P(Z_{i,j} = 1 | Z^l_{i,j} = 1)$ which is computed from a histogram of the training data for a particular (l^{th}) data source. This reflects the degree of confidence that biologists have come to associate with the interactions predicted from the l^{th} data source. The multiplication of posterior probabilities is equivalent to the addition of log-likelihoods of generation from each of the various data sources. The expression above decomposes the overall structure of the relationship into a product of marginal conditionals due to the assumed independence of the various data sources.

We now explore manifold embedding [2,4] as a method to incorporate the probability weights obtained from the interaction probability matrices to bring those genes closer which have a higher probability of interaction. For understanding transcriptional regulatory mechanisms, it can be hypothesized that the genes in close vicinity to a gene of interest are either co-regulated or potentially involved in the regulation of the target gene (through its product). A good embedding would use these diverse data sources to reflect such relationships.

2. LAPLACIAN EIGENMAP EMBEDDING

Suppose we are investigating the role of $(K - 1)$ genes in relation to our target gene (*Gata3*) - we proceed as follows:

- Standardize these K gene expression profiles to 0 mean and unit variance. Notice that the Euclidean distances become the Pearson correlation measure.
- Build the $K \times K$ dimensional weight matrix W from the Hadamard product of the L interaction probability $(P(Z^l_{i,j} = 1)^L_{l=1}$) matrices, from each of the L different sources of data.

1-4244-0385-5/06/$20.00 ©2006 IEEE

- Find n Nearest Neighbors using the Euclidean distance (or within some ϵ-neighborhood). Assign weight $W_{i,j} = p_{i,j}$, from (1) for the pair (i, j), for each of the $\binom{K}{2}$ gene pairs.
- Form the Graph Laplacian [2]:

$$L_{i,j} = \begin{cases} d_i = \sum_k W_{i,k} & \text{if } i = j; \\ -W_{i,j} & \text{if } i \text{ is connected to } j; \\ 0 & \text{otherwise.} \end{cases}$$

- Solve $min_y y^T Ly = \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 W_{i,j}$
- Embed the co-ordinates to a lower dimensional manifold, using the solution (the Laplacian Eigenmap) obtained from the minimization above.

3. DISCUSSION

We demonstrate the utility of the presented approach to understand the mechanisms underlying transcriptional regulation of the *Gata2/Gata3* genes in the developing kidney [3]. The primary source of data used to obtain distances is the microarray expression profiles of 47 genes known to be co-expressed with *Gata3* in the embryonic kidney. These are obtained from http://genet.chmcc.org. A large amount of data encompassing literature mining, microarrays, protein-protein interactions have been available from the STRING database (http://string.embl.de/) - for most of the $K = 48$ genes selected above, a lot of functional information from several experiments is available. For our purpose, we find the strength of association between any two genes i and j using significance scores from three different sources:

- Phylogenetic conservation of protein i's binding site in the upstream region of gene j.
- Interaction of Protein i with Protein j.
- Directed Information [3], measuring causality in expression of gene j due to gene i - based on microarray expression.

A common approach used for studying transcriptional regulatory mechanisms is by association. The hypothesis underlying this is that if genes are co-clustered/correlated, they are co-regulated, i.e. have a common set of controls. Since we are interested in the transcriptional regulatory mechanisms of *Gata3*, we look for genes which are in a ϵ-neighborhood of *Gata3*. From the embedded manifold in two-dimensions as shown in Fig.1, we observe that the *PPAR, Lamc2, Pax2* genes are among several which are 'in close proximity'(and possibly functionally relevant in transcriptional regulation) to the *Gata3* gene. This is interesting since each of these have phylogenetically conserved TF binding sites in the *Gata3* promoter. It is to be noted that *Gata3*'s family member *Gata2* is in the cluster on the top left, indicating that though it is expressed, the influence of the TF genes is more pronounced and hence, they are closer in the network. We note that this embedding has integrated information from three very different data sources to build this 'proximity map' of genes. These findings are currently being verified in the laboratory.

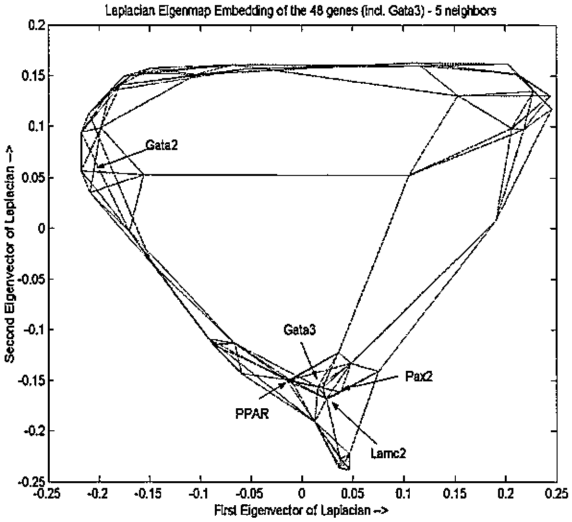

Fig. 1.

4. CONCLUSIONS

We have presented a methodology to understand the mechanisms underlying transcriptional regulation of a gene by combining various available data sources via a *modified Laplacian Eigenmap* technique. This framework provides a common ground both for the integration and visualization of diverse data sources for understanding physiological processes.

5. REFERENCES

[1] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, and Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in S. cerevisiae). Proc Natl Acad Sci USA 100(14): 8348-53, 2003.

[2] M. Belkin, P. Niyogi , Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, Neural Computation, June 2003; 15 (6):1373-1396.

[3] A.Rao, A.O. Hero, D.J. States, J.D. Engel, Inference of Biologically Relevant Regulatory networks using Directed Information, accepted to ICASSP 2006.

[4] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux and M. Ouimet. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps,and Spectral Clustering, In Advances in Neural Information Processing Systems, 2004.