Estimating Drivers Cell State Transitions using Gene Regulatory Network Models

Daniel Schlauch[1] and Kimberly Glass[2,3] John Quackenbush[1,3]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA 02215
[2]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 20015
[3]Department of Medicine, Harvard Medical School, Boston, MA 20015

November 23, 2015

Abstract

Abstract In the language of systems biology, the state of a cell can be represented by a gene regulatory network that characterizes the gene transcriptional processes that are active in that cell type. And transitions that occur in a wide range of biological processes, ranging from development to disease, can be thought of as transformation of the gene regulatory network from its initial state to its final state. Here we propose a regression-based generalization of the PANDA method for gene regulatory network inference for individual states, and a linear algebra approach to modeling cell state transitions, identifying transcription factors that alter the network structure as cell states change.

Modeling cell state transitions as a problem in gene regulatory network transition

Cell state transitions—such as those that occur during development, or as healthy tissue transforms into a disease phenotype—are fundamental properties of biological systems. Understanding what drives these transitions, and modeling the processes, is one of the great open challenges in modern biology. One way to conceptualize the state transition problem is to imagine that each phenotype has its own characteristic gene regulatory network, and that there are a set of processes that are either activated or inactivated to transform the

network in the initial state into that which characterizes the final state. Identifying those changes could, in principle, help us to understand not only the processes that drive the state change, but also how one might intervene to either promote or inhibit such a transition. The starting point for modeling cell state transitions is to model the initial and final cell states. One might imagine that the initial and final cell states consist of characteristic processes, some of which are shared (sometimes referred to as "housekeeping" functions) and others which are unique to the particular state. The way we understand these processes is that they are controlled by gene regulatory networks in which transcription factors (and other regulators) moderate the transcription of individual genes whose expression characterizes the state. One way to represent such processes is to draw a directed network graph, in which transcription factors and genes are nodes network in the network, and edges represent the regulatory interactions between transcription factors and their target genes that are active in, and characteristic of, a particular cellular state. One way of representing such a network, with interactions between m transcription factors and n target genes, is as a binary $m \times n$ "adjacency matrix," with 1's representing active transcription factor-target interactions, and 0's representing the lack of a transcription factor-target gene regulatory interaction. One can then think of a cell fate transition as the process that transforms the network in its initial state to its final state form, adding and deleting edges to remake the network that characterizes one phenotype into that which characterizes the other. Using the adjacency matrix formalism, one can think of this as a problem in linear algebra in which we attempt to find an $m \times m$ "transition matrix" $\mathbf{T}$, subject to a set of constraints that approximates the conversion from the initial network's adjacency matrix $\mathbf{A}$ into the final network's adjacency matrix $\mathbf{B}$, or

$$\mathbf{B} = \mathbf{AT}$$

While it is appealing to conceive of this process as deterministic, reflecting a change from one well-defined phenotype to another, in truth the situation is much more complex. Neither the initial nor the final phenotype is discrete, but each falls into a continuum of states, which, on average, captures the features of that phenotype. Indeed, within each tissue there are many, many cells, each of which is its own particular instance of that tissue—with unique patterns of gene expression and individual regulatory processes. In the language of adjacency matrices, what this means is that rather than representing each state by a matrix with binary entries, what one should do is use a representation in which entries are continuous, representing the strength of the transcription factor-target gene interaction averaged over the collection of samples (or cells) representing each state. And consequently, the problem of estimating the transition matrix is generalized to solving , where E is an $m \times n$ error matrix representing the uncertainty in the estimation of the individual edges. In this formalism, modeling the cell state transition is equivalent to estimating the appropriate transition matrix T that maps how the transcription factor-target gene interactions are "rewired" between states. And one could hypothesize that the drivers of the cell state transition are those transcription factors that are have the greatest change

in the targets that they regulate. In evaluating the state transitions, we recognize the limitations of current network inference methods to predict individual edgeweights. It's therefore of interest to combine measurements across sets of edgeweights in order to extract meaningful signal from a network perturbation. Effectively, we approached the problem as a dimension reduction problem with the goal of identifying high-influence systematic regulatory network alterations rather than isolated independent events. There are many existing methods for reducing a high-dimensional matrix such as a gene regulatory adjacency matrix. Commonly, Principal Components Analysis (PCA) identifies eigenvectors which can reconstruct the greatest degree of variance from the original data. One drawback of this approach is the lack of interpret-ability of these vectors. Our transition matrix approach can be considered as a data reduction method which (1) preserves the intuitive interpretation of its vectors and (2) utilizes our expectation that meaningful network transitions will occur via biologically systematic alterations and not via random, independent edge alterations.

BERE: A regression-based approach to modeling gene regulatory networks

In 2013, we described PANDA [2], a method [6] for estimating gene regulatory networks that uses "message passing" [1] to integrate multiple types of genomic data. PANDA begins with a prior regulatory network based on mapping transcription factor motifs to a reference genome and integrates other sources of data, such as protein-protein interaction and gene expression profiles, to estimate individual sample networks. While PANDA has proven to be very useful in a number of applications [3–5], its iterative approach to edge-weight optimization limits its utility in situations requiring a large number of network bootstrap estimations. To address this limitation, we developed BERE, Bipartite Edge Reconstruction from Expression. BERE approaches the network inference problem by considering the available evidence of an edge for each possible TF-gene pair. This evidence can be divided into two components, referred to here as direct and indirect. Consider the edge between a TF and a gene, referred to here as TF_i and g_j, respectively. The direct evidence, $d_{i,j}$, consists of the squared conditional correlation of the g_i and g_j given all other regulators of g_i. Where g_i is the gene which encodes TF_i

$$d_{i,j} = cor\left(g_i, g_j | \left\{g_{k,-j} : k \neq j, k \in \mathbf{TF}\right\}\right)^2$$

Naturally, the use of direct evidence inadequately captures regulatory relationships due to the impacts of technical noise and numerous biological external factors such as stable or transient protein-protein interactions, post-translational modifications, etc. which may confound or modify a regulatory effect. These sources of confounding and variability in the expression pattern of a gene coding a TF may obscure the effects it has on all of its target genes. Therefore it is of value if we can complement our estimate of the likelihood of a regulatory

mechanism by aggregating the information from the gene expression patterns of all suspected targets of transcription factors. PANDA achieves its superior performance in part by convergence towards "agreement", whereby large collections of gene expression patterns must agree with the proposed regulatory structure in order to claim an interaction. Similarly, BERE looks for agreement between the gene expression patterns of large sets of co-targeted genes. We refer to this feature as indirect evidence and can achieve this by again utilizing our set of regulatory priors. In this portion of the analysis we suspend the recognition of a TF as a member of the gene list and instead consider each of the m TFs to be binary classifications across the entire gene list. Class labels are determined by the presence or absence of a sequence binding motif for that TF in the vicinity of the gene.

The indirect evidence between the two nodes, $e_{i,j}$, represents the fitted probability that g_i belongs to the class of genes targeted by TF_j. g_i is considered to be a new observation placed into the $n-$dimensional space separated by transcription factor targets and non-targets. To divide up the space, BERE uses a regularized logistic regression on the gene expression data with the training set taken to be all genes and the training labels taken to be the existence or non-existence of a known sequence motif for TF_j upstream of g_i. The penalized model matrix comes from the recognition that correlations between co-regulated genes will be most strong when the TF_j is most prevalent. We therefore use the abundance of TF_j to weight the penalized model matrix, providing increased sensitivity for detecting coexpression for those samples in which we most expect it to occur. To build each of our classifiers we use the $L2$ regularization with the penalized model matrix, $\mathbf{Q}$, a diagonal matrix with weights equal to the the inverse expression value of the transcription factor. Effectively, we maximize the penalized logistic likelihood function

$$\sum_{i=1}^{n} log\left[exp\left(\beta'\mathbf{x_i}\right)^{Y_i}\left\{1-exp\left(\beta'\mathbf{x_i}\right)\right\}^{1-Y_i}\right]-\lambda\beta'\mathbf{Q}\beta$$

This computation is run using the R package "penalized", with the penalty term lambda estimated via default 5 fold cross validation.

By scoring each gene according to the strength of indirect evidence for a regulatory response to each of the TFs, we can combine this with the direct evidence of regulation (squared conditional correlation of expression for genei and TF_i). The appropriate manner in which to combine direct and indirect evidence remains an open question. Though both measures are bounded by [0,1] their interpretation is quite different. The direct evidence can be considered in terms of it's conditional gene expression R^2 between nodes, while the indirect evidence is interpreted as a probability. We use a non-parametric approach to combine evidence. The targets of each TF are then ranked and combined as a weighted sum, $w_i = (1-\alpha)\left[rank\left(d_i\right)\right] + \alpha\left[rank\left(e_i\right)\right], i \in \{1,\ldots,n\}$. Our choice of the weight, α, here is based on empirical evaluation, and perhaps not surprisingly, is loosely correlated with organism complexity. In validation sets from Yeast, the optimal alpha was observed near $\alpha = .9$ while simpler E. coli

datasets saw an optimal value of $\alpha = .6$ and an in silico dataset, optimality was achieved at $\alpha < .5$ This naturally reflects the fact that the increased complexity of the network necessitates the use of larger scale agreement between genes, rather than a reliance on pairwise correlations between potentially noisier and more complex expression patterns.

TM significantly improves TF-TF edge estimation from simulated gene expression data

To evaluate the ability of our method to recover edges between transcription factors, we generated simulated gene expression data. We began by generating a true adjacency matrix, $M_{0(p\times q)}$, describing the weighted edges between p genes and q transcription factors. A state transition was generated by sampling 100 TF-TF pairs and adjusting the edgeweight at the corresponding point on the true adjacency matrix. These TF-TF pairs ultimately represent the edges that we seek to recover with the size of the adjustments are the parameters of interest. We sampled from a multivariate Gaussian distribution with the off-diagonal of the variance-covariance matrix, Sigma, defined as the $M_0 M_0'$. Furthermore, we scaled the magnitude of the diagonal of Sigma to achieve the desired proportion of noise. This sampling represented our simulated control samples. The adjusted adjacency matrix, M_0, was similarly used to generate simulated expression data for the cases group. Next, we reconstructed the networks from our expression data using a set of commonly used network inference methods- Weighted Gene Correlation Network Analysis (WGCNA), Topological Overlap Map (TOM), Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE), Context Likelihood of Relatedness (CLR), Passing Attributes between Networks for Data Assimilation (PANDA) and simple Pearson correlation (PC). We applied the transition matrix with default parameters on each case-control pair of networks. For comparison, we estimated the difference from case to control in edgeweights derived from the direct edge prediction using each network inference method. The predictions for the TM approach and the direct approach were evaluated by the area-under-the-curve of the receiver-operator-characteristic (AUCROC) with the true transition adjustments taken as the gold standard. For each of the network inference methods tested, we found substantial improvement in the predicted transitions over the direct network inference method. In many cases, the edgeweight difference (row 2) was not statistically significantly better than chance at predicting transitions, but when the TM was applied (row 3) a strong predictive signal appeared. In other cases, an existing signal was observed using the direct approach, but was significantly improved with the application of the TM. Simulation table goes here The intuition behind the improvement is simple. While the estimation of a TF-TF edge is typically evaluated via some pairwise gene expression pattern which may be rife with technical and biological noise, the TM approach borrows information from all downstream targets in estimating the relative change

in relationship between the TFs.

Differential transcription factor involvement

Many mechanisms which may be differentially present, such as RNA degredation, post-translational modification, protein-level interactions and epigenetic alterations have the ability to impact downstream targeting without impacting the expression level of the TF itself. It may be of particular scientific or therapeutic interest to identify those TFs which have undergone significant overall changes in behavior between controls and cases. With that objective in mind, we express the statistic- differential Transcription Factor Involvement (DTFI), as a measure for quantifying this property.

$$s_j = \frac{\sum_{i=1}^{m} I\left(i \neq j\right) \tau_{i,j}^2}{\sum_{i=1}^{m} \tau_{i,j}^2}$$

DTFI can be loosely interpreted as the proportion of TF targeting patterns which is explained by the targeting patterns of other available TFs. This measure, a statistic on the interval $[0, 1]$ seeks to elucidate transitions which are systematic, informative, and non-arbitrary in nature by capturing only the edgeweight signal for which there is an attributable regulatory pattern. The distribution of this statistic under the null has a mean and standard deviation which depend on the motif structure. In particular, both mean and standard deviation are increased for TFs which have fewer prior regulatory targets. From a statistical perspective, TFs with relatively more targets are able to generate more stable targeted expression patterns, which leads to more consistent estimates in "agreement" algorithms such as PANDA and BERE. From a biological perspective, increased motif presence may indicate that the TFs are more likely to be ubiquitous housekeeping proteins that do not meaningfully alter their involvement between cases and controls. The dependence of the null distribution on the motif structure is addressed via the following resampling procedure.

1. Gene expression samples are randomly assigned to case and control forming the null-case and null-control with group sizes preserved.
2. GRNs are reconstructed for the null-case and null-control with the same prior regulatory structure.
3. The transition matrix algorithm is applied for the two null networks.
4. The differential TFI is calculated for each TF.
5. Repeat 1-4 1000 times.

Transition Matrix finds concordance in independent datasets for COPD

We applied our method to three case-control datasets for Chronic Obstructive Pulmonary Disease (COPD)- Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE), the COPDGene study, and Lung Genomics Research Consortium (LGRC). Each of these studies consisted of gene expression assays from blood samples of COPD patients and smoker controls. We separately applied our BERE network inference approach on cases and controls and computed the transition matrix. Top significance hits for DTFI showed strong concordance between each of the datasets. [Results table here?] Two of the top 3 hits, NRF1 and GABPA have been implicated in a mitochondrial mechanism for disease progression [has this been published?] Interestingly, the majority of the TFs identified as differentially involved do not exhibit significant differential gene expression. This suggests that for these proteins, their role in the disease may not occur until the post-transcription stage. It also suggests that conventional gene expression analysis is insufficient for identifying many of the TF drivers of disease.

In solving for the transition matrix T, what we are attempting

One of the fundamental problems is biology is modeling the transition between biological states such as that which occurs during development or as a healthy tissue transforms into a disease state. While it is appealing to conceive of this process as deterministic, reflecting a change from one well-defined phenotype to another, in truth the situation is much more complex. Neither the initial nor the final phenotype is discrete, but each falls into a continuum of states, which, on average, captures features of that phenotype. Indeed, within each tissue there are many, many cells, each of which is its own particular instance of that tissue—with unique patterns of gene expression and individual regulatory processes. The same is true when considering individuals as each healthy and disease state is unique to each member of a study population. One way to conceptualize the state transition problem is to imagine that each phenotype has a characteristic gene regulatory network and that there are a set of processes that are either activated or inactivated to transform the network in the initial state into that characterizing the final state. Identifying those changes could, in principle, help us to understand not only the processes that drive the state change, but also how one might intervene to either promote or inhibit such a transition.

The problem of gene regulatory network inference

As our ability to generate large-scale, integrative multi-omic datasets has grown, there has been an increased interest in using those data to infer gene regulatory networks to model fundamental biological processes. While there have been many network inference methods published, each of which uses a different approach to estimating the "strength" of interactions between genes (or between

transcription factors and their targets), they all suffer from the same fundamental limitation. Every method relies on estimating weights that represent the likelihood of an interaction between two genes and then setting a threshold to identify "real" (high confidence) edges. While setting edge confidence thresholds allows us to graphically represent networks and allows us to compare networks based on the presence or absence of edges, it ultimately requires that we discard information regarding those "weak" edges that fail to reach significance. Now one could argue that discarding low significance edges is sensible as one common goal in network inference is to deduce a single, high confidence network model that represents a particular phenotype under study or, in some cases, a transition between phenotypes. An alternative view is that these

References

[1] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[2] Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5):e64832, 2013.

[3] Kimberly Glass, John Quackenbush, Edwin K Silverman, Bartolome Celli, Stephen I Rennard, Guo-Cheng Yuan, and Dawn L DeMeo. Sexually-dimorphic targeting of functionally-related genes in copd. *BMC systems biology*, 8(1):118, 2014.

[4] Kimberly Glass, John Quackenbush, Dimitrios Spentzos, Benjamin Haibe-Kains, and Guo-Cheng Yuan. A network model for angiogenesis in ovarian cancer. *BMC bioinformatics*, 16(1):115, 2015.

[5] Taotao Lao, Kimberly Glass, Weiliang Qiu, Francesca Polverino, Kushagra Gupta, Jarrett Morrow, John Dominic Mancini, Linh Vuong, Mark A Perrella, Craig P Hersh, et al. Genome medicine. 2015.

[6] Catharina Olsen, Kathleen Fleming, Niall Prendergast, Renee Rubio, Frank Emmert-Streib, Gianluca Bontempi, Benjamin Haibe-Kains, and John Quackenbush. Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics*, 103(5):329–336, 2014.