Identifying structural changes in case-control gene regulatory networks, with applications in COPD

Dan Schlauch[1,3], Kimberly Glass[2], John Quackenbush[1,3]

[1]Department of Biostatistics, Harvard School of Public Health; [2]The Channing Division of Network Medicine, Brigham and Women's Hospital;
[3]Department of Computational Biology and Biostatistics, Dana-Farber Cancer Institute

Introduction

The development of high-throughput technologies over the last two decades has brought significant promise towards understanding biology and shed light on the progression of human disease, but understanding the networks which drive these changes remains a challenge. Often we are tasked with identifying regulatory networks based off of static information, such as the bibliome, and a set of experimental assays taken at the mRNA level. However, we know that many regulatory mechanisms, such as methylation, histone modification, protein interaction or degradation, etc. occur in the space between adjacent gene expression measurements.

Our goals can be divided into two distinct parts, (1) the construction of gene regulatory networks and (2) the analysis of the structural changes between those networks. Due to the complexity of the underlying networks and the high dimensionality of typical datasets, these challenges remain open problems. In our work we develop (1) a novel method for GRN inference and (2) a technique developed for gaining meaningful insight into network transformations between cases and controls in a complex disease.

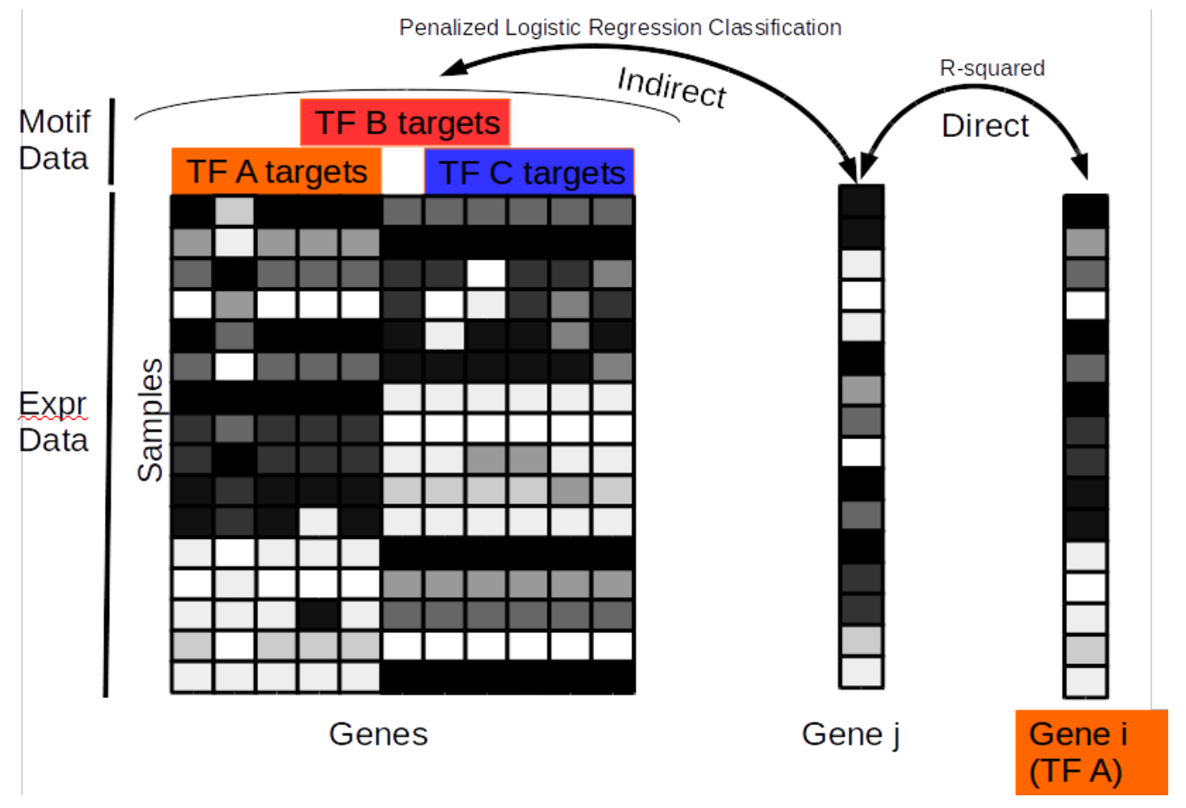
Methods

Gene Regulatory Network Reconstruction

Consider a gene expression matrix consisting of *p* rows of genes across *n* samples. Additionally, consider a set of regulatory priors consisting of a mapping between *m* known TFs and the variable set of genes for which they are suspected of targeting. Our method uses a bipartite network modeling framework. We have developed a novel approach, **B**ipartite **E**dge **R**econstruction from **E**xpression data (BERE, pronounced "bear"), for inferring transcription factor to gene interactions.

BERE operates by dividing the the evidence of regulation into 2 parts.
1.) *Direct* evidence, measured as the coexpression of a transcription factor with a gene.
2.) *Indirect* evidence, measured as the agreement in expression between other targets of the TF.

Indirect evidence is measured via a penalized logistic regression model, with the penalty Model matrix defined as the inverse expression of of the TF.
Combining the ranks of direct and indirect evidence yields a BERE score which is shown to predict TF-DNA binding, as measured by area under ROC curve against ChIP-chip results, with superior performance to existing methods.



Network Transition Analysis

Consider two adjacency matrices, **A**, **B** representing the two GRNs estimated from a case-control study. Each matrix has dimensions *p* x *m* representing the set of *p* genes targeted by *m* TFs. We seek a matrix, **T**, such that **B** = **AT** + **E**. In other words, we formulate the network transition as a regression problem where we are seeking to find a linear combination of TF targeting patterns in the control group which predict the targeting patterns in the cases group. This allows for the natural interpretation of the transition matrix **T**, as representing the transfer of attributes from one TF to another as we go from control to cases.

The transition matrix is computed via a series of L_1 regularized regressions, using the following form for the penalty model matrix and error function.

$$\mathbf{Q}_{i,j} = \begin{cases} 1 & for\ i = j \neq k \\ 0 & elsewhere \end{cases} \qquad \sum_{i=1}^{p}\left(\mathbf{B}_{i,k} - \sum_{j=1}^{m} A_{i,j}\mathbf{T}_{j,k}\right)^2 + \lambda\sqrt{\beta'\mathbf{Q}\beta}$$

The transition matrix can then be evaluated for biologically relevant targeting transfers. Differential TF involvement is measured by the scaled sum of squared off-diagonal mass for each column, yielding a statistic which can be loosely interpreted as the proportion in variability in TF-targeting which is explained by variability in TF-targeting in the control network. Additionally, off diagonal mass can be treated as an indication of a general interaction which involves the TF corresponding to the row and column of the matrix entry.

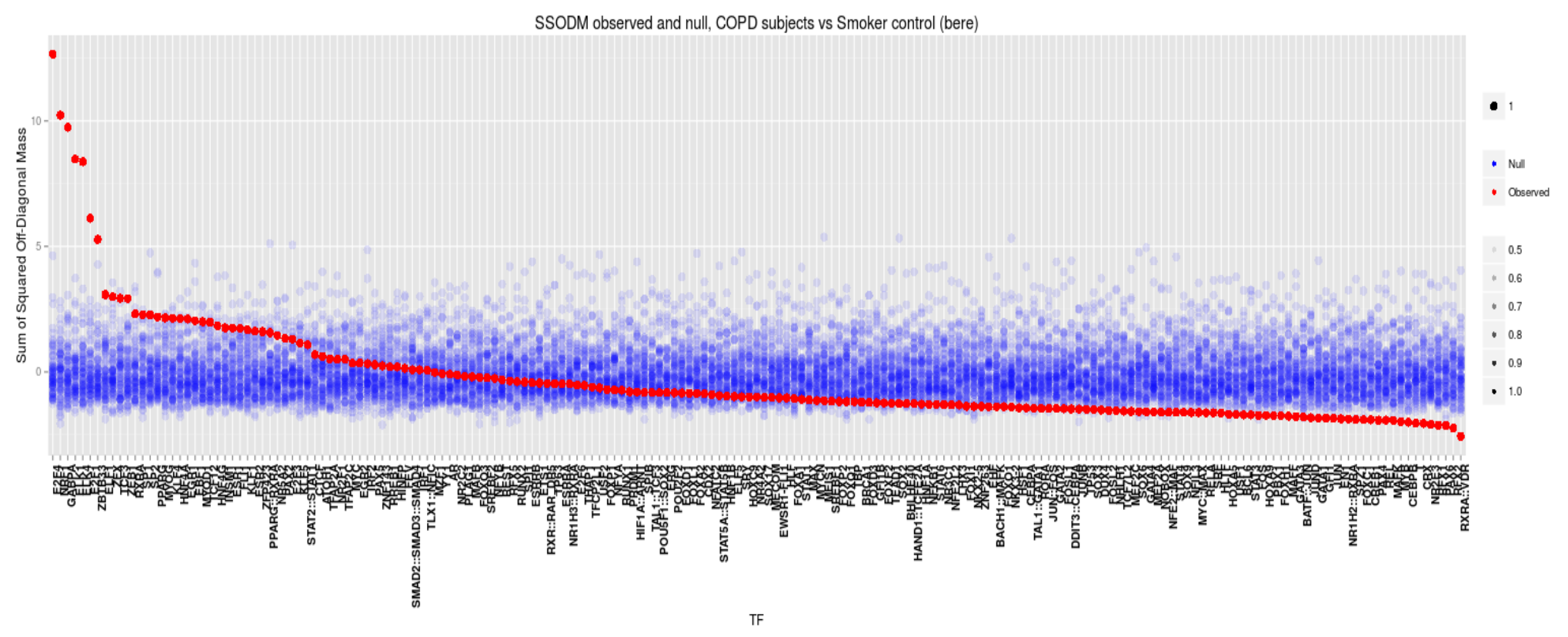
Because the distribution in off-diagonal mass is dependent on numerous features beyond the gene expression data, we estimate the distribution of these statistics based on the null hypothesis of identical regulatory network structure by using a permutation base resampling approach. By randomly reassigning samples to case and control we can generate a large number of pairs of GRNs for which we can estimate the transition. In this manner, we can obtain the degree of statistical significance for each TF involvement or TF interaction.

Results

To test our methods, we used four test datasets of increasing biological complexity- (1) in silico, (2) E. coli, (3) Yeast, and (4) human. Data from the first three sources was obtained from the publicly available DREAM5 challenge. This challenge asked contestants to infer gene networks from expression data alone, using a gold standard for evaluation.

BERE had comparable or superior performance to existing methods such as PANDA, SEREND, WGCNA in predicting regulatory structure while running in a fraction of the time.

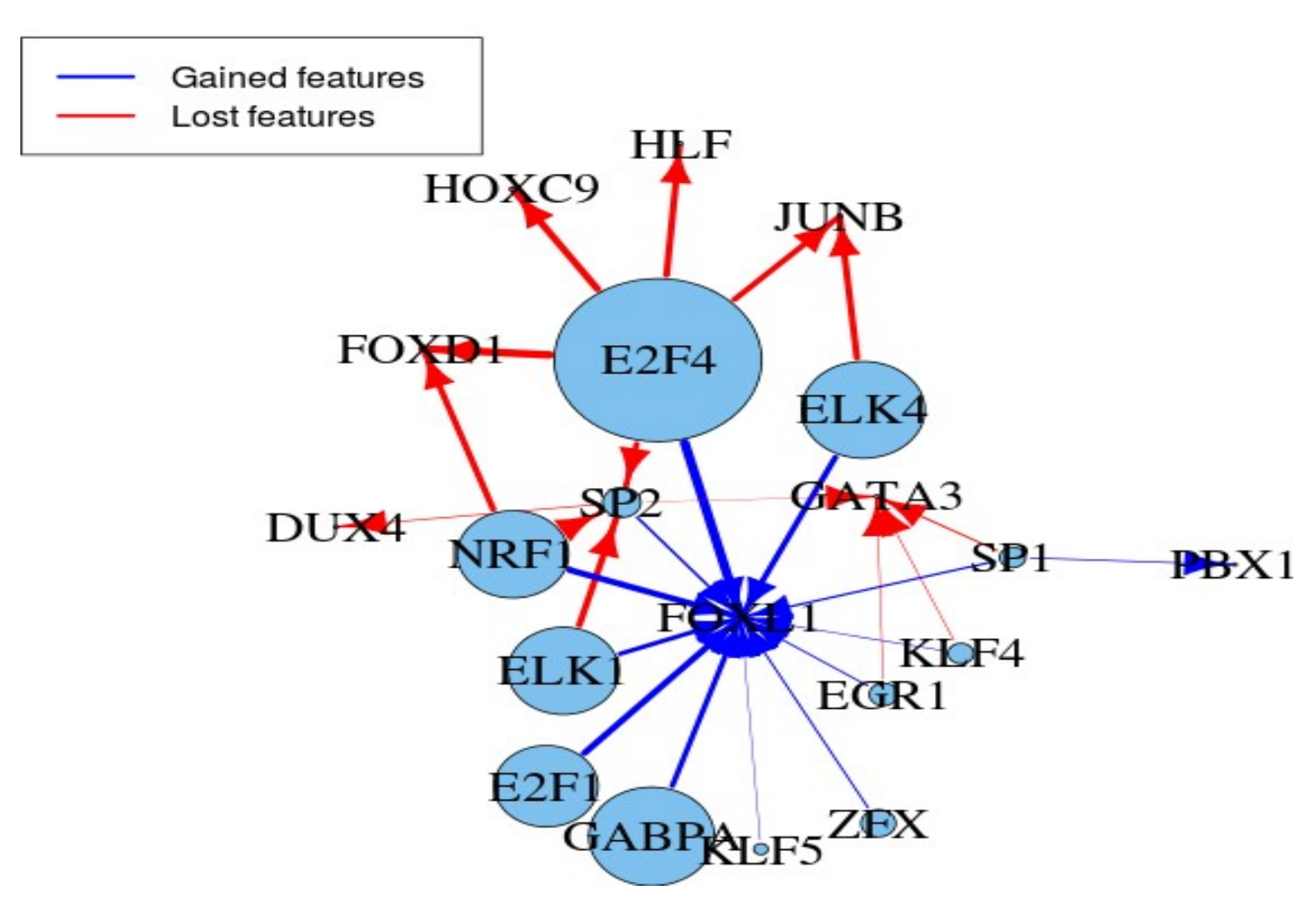
We applied the BERE method on data from the ECLIPSE study on COPD- consisting of a set of 136 COPD cases and 84 controls. We investigated the changes which characterize the transition from control to COPD and discovered several significantly differentially involved transcription factors. The majority of these significant TFs were not differentially expressed between cases and controls, suggesting a post-translational mechanism for disease involving these proteins.



Differential Transcription Factor Involvement (Top transcription factors COPD vs Control)					
	Score	p-values	FDR	LIMMA	Notes
E2F4	12.666	0.0000	0.0000	0.337	Binds EGR-1, SMAD3. Tumor suppression.
NRF1	10.232	0.0000	0.0000	0.215	Acts on nuclear genes encoding respiratory subunits and components of the mitochondrial transcription and replication machinery.
GABPA	9.747	0.0000	0.0000	0.816	Related to NRF1, involved in activation of cytochrome oxidase expression and nuclear control of mitochondrial function
ELK1	8.480	0.0000	0.0000	0.080	Binds to the the serum response factor
ELK4	8.379	0.0000	0.0000	0.000	Binds promoter of the c-fos proto-oncogene
E2F1	6.126	0.0000	0.0000	0.714	E2F family...
ZBTB33	5.281	0.0000	0.0000	0.602	shown to interact with HDAC3, Nuclear receptor co-repressor 1
ELF1	3.083	0.0010	0.0242	0.301	primarily expressed in lymphoid cells
ZFX	2.998	0.0014	0.0285	0.987	gene on the X chromosome

We further investigated the specific alterations in targeting which occurred in the transition and identified several meaningful "targeting transfers" which occurred among significant TFs.
The graph on the right indicates that patterns are emerging with respect to regulatory targeting. E2F4, the most significantly differentially involved TF loses the targets characterized by several other TFs, suggesting an alteration which impacts its ability to regulate these pathways in COPD. It is noteworthy that E2F4 and other differentially involved TFs do not exhibit significant differential expression, suggesting that these changes in behavior occur at the post-translational level.

Transition: SMC -> COPD



Conclusions

The discovery of regulatory network changes in case-control studies is an area of significant interest. The ability to infer changes in the behavior of the TFs which control regulation has widespread implications in drug targeting, predicting response to therapies, identifying disease subtypes and identifying driver mechanisms for disease. Here we developed a novel gene regulatory network reconstruction algorithm and network transition method which address the problems. Our network inference method is shown to provide superior regulatory predictions compared to several existing methods and has a dramatically shorter execution time. This promotes the use of resampling algorithms, such as the one performed here in our transition method. The transition method proposed here provides an intuitive dimension reduction framework which allows us to interpret the results in a biologically meaningful manner.
Together, these two approaches have been applied to discover several TFs for which their targeting pattern significantly characterize the case-control transition in COPD. Furthermore, this approach allows us to specifically identify the manner in which these TFs have altered their targeting.

VE RI TAS
BWH
DANA-FARBER
CANCER INSTITUTE