A novel supervised learning algorithm for detection of cis-regulatory modules in promoters and enhancers of ES cells

Matthew Ziegler, advised by Mark Craven (Biostatistics and Medical Informatics, Computer Science), Ron Stewart (Morgridge Institute), and Scott Swanson (Morgridge Institute)

Abstract

The identification of regulatory relationships is an important part of genetics and developmental biology research. We propose to develop a novel supervised learning algorithm, modeled after the multiple instance learning problem, to try to identify the regulatory mechanisms at work during BMP4-induced ES cell differentiation. Pluripotent ES cells differentiate into a poorly-understood "mesoendoderm" state when treated with BMP4. To identify genes with expression affected by BMP4 treatment, we will use a data set of gene expression measurements collected from ES cells before and after BMP4 treatment. We will use histone mark patterns to predict enhancer and promoter locations for the genes in our experiment, and then identify possible transcription factor binding sites by scanning our set of promoters and enhancers against a set of known motifs. We will then develop a novel supervised learning algorithm to learn a model of binding sites in enhancers and promoters that separates BMP4-affected genes from unaffected genes to identify the transcription factors of interest in this system. Because of the recent availability of enhancer datasets, our algorithm will be applicable to similar regulatory module finding tasks.

Introduction and Literature Review

Embryonic Stem (ES) cells are pluripotent cells which have potential to differentiate into any cell type in the body. ES cell research is a key component of understanding animal development, and may lead to a wide array of new regenerative therapies. An important task in ES cell research is understanding how stem cells are "programmed" - how cells are signaled to differentiate into different cell types.

When treated with Bone Morphogenetic Protein 4 (BMP4), pluripotent ES cells have been shown to differentiate into an intermediate state between the mesoderm and endoderm germ layers, where some cells will continue to differentiate into mesoderm and others into endoderm (Yu et al., 2011). This "mesoendoderm" state is not well understood. To learn more about this stage of embryo development, we are interested in identifying the regulatory mechanism behind the changes in gene expression during differentiation.

To identify genes which appear to be affected by BMP4 treatment, a recent experiment by the Morgridge Institute's Center for Regenerative Biology treated ES cells with BMP4, and gene expression in the cells was measured before and after treatment. To search for a common regulatory mechanism shared by the affected genes, we can identify promoters and enhancers for all of the genes in the experiment, and then then find a combination of transcription factor (TF) binding sites in the promoters and enhancers that separates the affected

genes from the unaffected genes.

Recent work has shown that promoters and enhancers can be identified by examining chromatin histone marks (Mikkelsen et al., 2007; Heintzman et al., 2009). Histone marks are modifications of chemical groups on histone proteins, thought to be for regulatory purposes. Certain patterns of histone marks are associated with regulatory regions, eg. monomethylation of H3K4 in enhancers and trimethylation of H3K4 in promoters (Heintzman et al., 2007). Analysis of chromatin marks can also be used to infer which regulatory regions act upon which genes (Ernst et al., 2011).

To infer the regulatory mechanism (called a cis-regulatory module or CRM in the literature,) we will set up our task as a supervised machine learning problem. A supervised learning algorithm is an algorithm that, given a set of input and a corresponding set of output, tries to learn a function that maps the each input to its corresponding output. In our case, the sequences of the enhancers and promoters of each gene is the input, and the regulation or non-regulation of each gene is the output.

Many computational approaches to predicting regulatory mechanisms (called cis-regulatory modules or CRM's) rely on supervised machine learning algorithms to construct models which separate a set of genes of interest (eg. genes thought to be regulated together) from a set of "background" genes (Elemento et al., 2007; Noto and Craven, 2007). Potential binding sites are predicted by "motifs" - short reoccurring sequences to which TF's are known to bind - in the genome sequence around the gene. The learning algorithm identifies a set of rules describing the motif occurrences that separate the genes of interest from the background genes. Some algorithms have been extended to include other information, such as spacial relationships between the motifs (Noto and Craven, 2007), or histone marks near the genes (Won et al., 2009). Our algorithm will utilize information about promoter and enhancers obtained from the histone marks.

Our task resembles the well-studied multiple instance learning problem, in which the learner is given a set of positive and negative "bags." Each bag contains several instances which are either positive or negative. The bag is positive if it contains at least one positive instance, and is negative otherwise. The learner knows which bags are positive, but not which instances are positive, and must learn a set of rules which uses information about the instances to separate the positive bags from the negative bags (Dietterich, 1997). In our case, the bags represent genes, and the instances represent promoters and enhancers. The learner must find a set

of rules regarding the motifs in each instance to separate the regulated genes from the unregulated genes. The model is illustrated in Figure 1.

Our task differs from multiple instance learning in two main ways. First, enhancers can act upon multiple genes, but in the multiple instance problem each instance belongs to exactly one bag. Second, in the multiple instance problem each instance is independent (is always either positive or negative by itself,) but we want to allow for the possibility of interaction between transcription factors in different regulatory regions.

To address these differences, we will create a novel algorithm to fit our task, extending the multiple instance problem. The objective of the algorithm will be to learn a set of rules that use the motifs in promoters and enhancers to separate out genes genes of interest (genes affected by BMP4 treatment) from background genes. Because of the recent availability of promoters and enhancer datasets, our algorithm will be applicable to other CRM-finding problems as well.

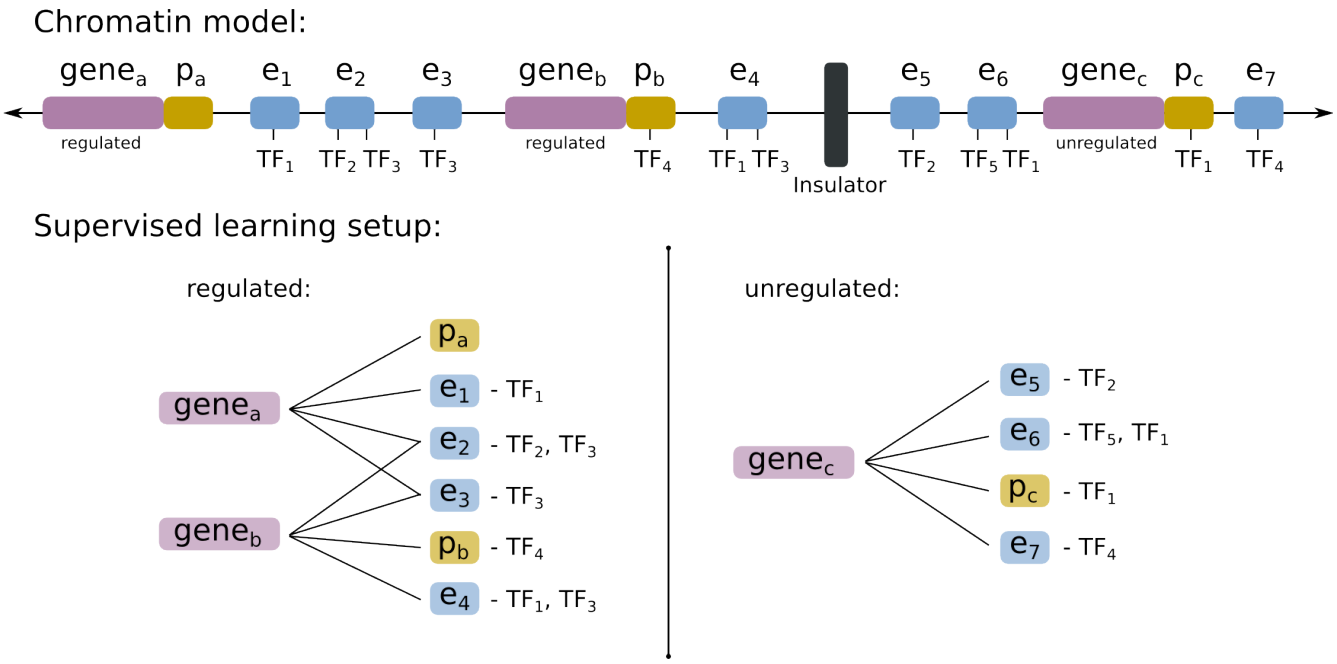

Figure 1: Chromatin model and supervised learning setup. Promoters are denoted with "p" and enhancers with "e." In our supervised learning problem, genes are represented as "bags of enhancers and promoters," and enhancers and promoters are represented as "bags of transcription factors." Genes are linked to enhancers by insulator boundries and chromatin state correlation across cell types. The task of the learner is to find a rule that uses the information about TF's and enhancers and promoters to separate the regulated genes from the unregulated genes.

Methods

Our approach to identifying the regulatory mechanisms active in BMP4-induced differentiation will consist of several consecutive steps. BMP4-induced gene expression data has already been collected. We will identify promoters and enhancers in our dataset, and then find TF binding sites in the promoters and enhancers. As the final step, we will develop a supervised learning algorithm to attempt to infer the regulatory mechanism.

To identify genes with expression affected by BMP4 treatment, 4-day-old ES cells were treated with BMP4 for 3 days. Gene expression was measured using microarrays both before and after treatment. To determine which genes had their expression upregulated, downregulated, or not regulated, we will compare the two expression measurements and choose thresholds that separate the groups.

We will identify promoters and enhancers for the genes in the experiment by using probabilistic models over chromatin immunoprecipitation data for histone proteins across the genome (Heintzman et al., 2009). Regulatory elements will be related to their target genes using insulator positions in the genome - insulator regions block enhancer action on genes (Burgess-Beusse et al., 2002). Comparison of chromatin states across different cell types can also be used to correlate enhancers with their respective genes (Ernst et al, 2011). New enhancer datasets are being released very frequently, so we may decide to use a different approach at the time of our experiment.

To locate TF binding sites, we will use a database of known motifs such as TRANSFAC (Matys et al., 2003). Our set of promoters and enhancers will be scanned against the known motifs, and a p-value will be computed for each possible motif occurrence (Touzet and Varré, 2007). We will determine a threshold for p-values for inclusion of motifs in our dataset.

We will set up our supervised learning problem similar to a multiple instance problem (Dietterich, 1997). Genes will be represented as bags of instances, and enhancers and promoters will be represented as instances. It will be the learner's job to learn a model that can separate regulated genes from unregulated genes.

We will extend a multiple instance learning algorithm to create a novel algorithm that fits our task. There are many multiple instance learning algorithms in use, and we will experiment with extending different algorithms to find one that best suits our task.

Timeline

1. Gene expression data collection is already completed
2. Enhancer and promoter prediction from histone data: June – July 2012
3. Transcription factor binding site prediction: August 2012
4. Supervised learning algorithm development: September 2012 – December 2012
5. Results and manuscript preparation: January 2013 – April 2013

References

1. Burgess-Beusse B, Farrell C, Gaszner M, et al. The insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci USA. 2002;99(Suppl 4):16433–7.
2. Dietterich, T. "Solving the Multiple Instance Problem with Axis-parallel Rectangles." Artificial Intelligence 89.1-2 (1997): 31-71.
3. Elemento, Olivier, Noam Slonim, and Saeed Tavazoie. "A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types." Molecular Cell 28.2 (2007): 337-50.
4. Ernst, Jason, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, and Lucas D. Ward. "Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types." Nature 473.7345 (2011): 43-49.
5. Heintzman, Nathaniel D., Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W. Ching, R. David Hawkins, Leah O. Barrera, Sara Van Calcar, Chunxu Qu, Keith A. Ching, Wei Wang, Zhiping Weng, Roland D. Green, Gregory E. Crawford, and Bing Ren. "Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome." Nature Genetics 39.3 (2007): 311-18.
6. Heintzman, Nathaniel D., Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, Leonard K. Lee, Rhona K. Stuart, Christina W. Ching, Keith A. Ching, Jessica E. Antosiewicz-Bourget, Hui Liu, Xinmin Zhang, Roland D. Green, Victor V. Lobanenkov, Ron Stewart, James A. Thomson, Gregory E. Crawford, Manolis Kellis, and Bing Ren. "Histone Modifications at Human Enhancers Reflect Global Cell-type-specific Gene Expression." Nature 459.7243 (2009): 108-12.
7. Matys, V., E. Fricke, R. Geffers, E. Gößling, and M. Haubrock. "TRANSFAC: Transcriptional Regulation, from Patterns to Profiles." Nucleic Acids Research 31.1 (2003): 374-78.
8. Mikkelsen, Tarjei S., Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P. Koche, William Lee, Eric Mendenhall, Aisling O'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S. Lander, and Bradley E. Bernstein. "Genome-wide Maps of Chromatin State in Pluripotent and Lineage-committed Cells." Nature 448.7153 (2007): 553-60.
9. Noto, Keith and M. Craven. Learning probabilistic models of cis-regulatory modules that represent logical and spatial aspects. Bioinformatics 23(2):e156-e162, 2007.
10. Touzet, Hélène, and Jean-Stéphane Varré. "Efficient and Accurate P-value Computation for Position Weight Matrices." Algorithms for Molecular Biology 2.1 (2007): 15. Print.
11. Won, Kyoung-Jae, Saurabh Agarwal, Li Shen, Robert Shoemaker, Bing Ren, and Wei Wang. "An Integrated Approach to Identifying Cis-Regulatory Modules in the Human Genome." Ed. Raya Khanin. PLoS ONE 4.5 (2009): E5501.
12. Yu, Pengzhi, Guangjin Pan, Junying Yu, and James A. Thomson. "FGF2 Sustains NANOG and Switches the Outcome of BMP4-Induced Human Embryonic Stem Cell Differentiation." Cell Stem Cell 8.3 (2011): 326-34.