Project Proposal: A Causal Approach for In-Silico Mutagenesis

Daniel Cizin, Stephen Malina

November 7, 2019

1 Proposal

1.1 Abstract

We hope to use a variant of Mendelian Randomization (MR) combined with a convolutional neural network (CNN) trained to predict transcription factor binding from sequence to identify whether pioneer transcription factors gate the binding of secondary transcription factors and affect chromatin accessibilty as hypothesized. In the process, we hope to validate that MR can work when applied to a graphical model in which relationships between variables are represented by neural networks. This work can potentially lay a foundation for using MR at scale to test hypotheses about complex network

1.2 Biological Motivation

Pioneer transcription factors are transcription factors whose binding plays a large role in opening up closed regions of chromatin. These transcription factors thus influence a wide range of important biological phenomena (Lai et al. (2018)). As a result, being able to verify causal relationships between both pioneer and secondary TF binding and pioneer TF binding and chromatin accessibility *in silico* would be valuable.

1.3 ML Methods

Our project will combine deep learning and causal inference. To predict transcription factor binding, we'll train convolutional neural nets to classify sequences with binary labels. To test the hypothesis that pioneer TF binding gates secondary TF binding, we'll adapt Mendelian Randomization, an instrumental variable (IV) method traditionally used by epidemiologists and biostatisticians, to verifying causal effects.

1.3.1 Mendelian Randomization & In-silico Mutagenesis

Classical Mendelian Randomization leverages the fact that genetic recombination randomizes gene assignments to estimate the causal effect of a phenotype on some disease (Didelez and Sheehan (2007)). The graphical model depicted by figure 1 represents the assumptions MR studies make. In a typical MR study, X represents some phenotype or treatment of interest, Y the outcome of interest, Z an allele believed to influence X, and U unmeasured confounding correlated with both X and Y. As captured by the arrow structure of the graph, in order for Z to be a valid IV, it must influence X directly, not influence Y outside of its influence on X and not correlate with U. Further discussion of the standard IV and MR assumptions can be found in Didelez and Sheehan (2007).

In place of natural genetic variation, we'll leverage in-silico mutagenesis and the presumed ability of neural networks to predict TF binding affinities for unseen variants of sequences found in the training data to probe the relationship between pioneer and secondary TF binding. In this setting, sequences and mutated versions of them will function as instrumental variables (Z in the diagram), binding of alleged pioneer TFs to these sequences as the phenotype or exposure (X in the diagram), and binding of alleged secondary TFs to these same

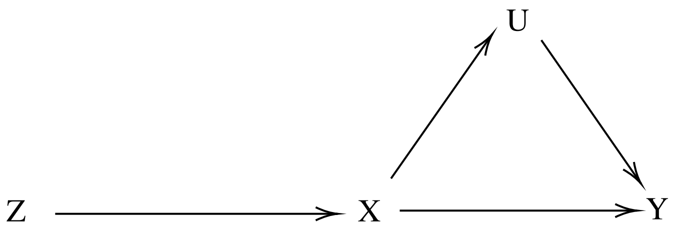

Figure 1: Graphical model depicting the standard Mendelian Randomization setting. The structure of the arrows reflects our hypothesis about the causal structure of our phenomenon and also the assumptions required to leverage IV approaches.

sequences as the outcome. The confounding variable, U, will represent the impact of chromatin accessibility and other unmeasured cell state on binding.

Similar to Zhao et al. (2018), our setting differs from classical MR in how we think about the relationship of our instrumental variable to our phenotype and outcome. In classical MR, we think of all of our samples as reflecting a single underlying instrumental variable, phenotype, and outcome set triad. For example, to study the impact of cholesterol levels on heart disease using some genetic variant, we could take measurements from many individuals with different cholesterol levels, genetic variants, and heart disease outcomes. In our case, we'll instead view each sequence and its variant as a different independent instrumental variable as acting on our phenotype. As part of our project, we'll formalize these assumptions and their implications.

1.4 Data Sources

Similar to Alipanahi et al. (2015) (DeepBind) and Zhou and Troyanskaya (2015) (DeepSEA), we intend to use a subset of ENCODE's *in vivo* transcription factor binding (ChIP-seq) and chromatin accessibility (DNase-seq) data to train and test our model. Initially, we intend to look at the FOXA1 and CTCF pioneer TFs and their corresponding secondary TFs, but hope to expand our model to more TFs once we have causal mutagenesis working with the simplified model.

1.5 Other Relevant Papers

1.5.1 TF Binding Prediction

Alipanahi et al. (2015) and Zhou and Troyanskaya (2015) provide benchmarks for validating our base CNN models' accuracy.

1.5.2 Causal Inference and Mendelian Randomization

Burgess et al. (2014) extend Mendelian Randomization to the scenario in which an instrumental variable may operate both directly and indirectly upon an outcome variable. Were we to find that all genetic variants we inspect impact binding affinity of both the pioneer and secondary TF, this paper's approach could prove useful for relaxing our assumption of no indirect effects.

1.6 Expected Contributions

We expect our contributions to be roughly equivalent. To the degree there will be specialization, it will be based off of Daniel's slightly deeper understanding of the biological side and Stephen's slightly deeper understading of causal inference.

1.7 Deliverables

1.7.1 Evaluation

Initially, we can evaluate the efficacy of MR for pioneer/secondary TFs by training a CNN on positive and negative pairs of example TFs and confirming that MR estimates a directionally correct causal effect in both cases. Assuming this initial analysis shows that MR seems to be doing something reasonable, we can train a larger model (similar to DeepSEA)

1.7.2 Concrete Goals

By the midcourse report deadline we hope to:

- Formalize our variant of MR, its assumptions, and how (if at all) they differ from Zhao et al. (2018) and Burgess et al. (2014).
- Train & test a CNN model for predicting TF binding for two TFs, one pioneer, one secondary.
- Write code to do find IV candidates and run MR analysis "on" CNN for TF binding.
- (Stretch) Also do a similar analysis for chromatin accessibility.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831.

Burgess, S., Daniel, R. M., Butterworth, A. S., Thompson, S. G., and Consortium, E.-I. (2014). Network mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International journal of epidemiology*, 44(2):484–495.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.*, 16(4):309–330.

Lai, X., Verhage, L., Hugouvieux, V., and Zubieta, C. (2018). Pioneer factors in animals and Plants-Colonizing chromatin for gene regulation. *Molecules*, 23(8).

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2018). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931.