Determining Causal Effects of Transcription Factor Binding on Chromatin Accessibility with Mendelian Randomization

Daniel Cizin
dmc2236@columbia.edu

Stephen Malina
sdm2181@columbia.edu

Abstract

Deep learning has proven to be a powerful tool in genomics for predicting biological phenomena. In this paper we use two convolutional neural networks in order to predict transcription factor binding of CTCF and chromatin accessibility. With these models in hand, we set up a Mendelian Randomization study in which we use in-silico mutagenesis as a means of generating instrument variables. This set up should enable us to demonstrate that CTCF binding has a causal effect on chromatin accessibilty. In doing so, we hope to provide a generic method for validating causal effects with data generated by neural networks that future work can build on.

1 Introduction

In recent years, many researchers have applied deep learning to achieve impressive results at predicting transcriptomic features such as transcription factor (TF) binding [1, 13], chromatin accessibility [6, 13], RNA binding protein (RBP) binding [12, 10, 7], and DNA methylation states [2] from sequence and sometimes other auxiliary features. These models' ability to make reliable predictions of transcriptomic features on never-before-seen sequences may allow them to one day guide future investigation while saving valuable and scarce experimenter time. However, little work has been done to try and validate hypothesized causal relationships between phenomena using predictions from these models.

While little of this work has been applied to predictions from neural networks, over the past century or so, researchers have made impressive strides towards defining the philosophical and algorithmic foundation for describing and testing for causal effects from observational data [9]. Mendelian Randomization (MR), one such causal inference technique, takes advantage of Nature's mostly random assignment of genetic variants to individuals to estimate the causal effect of phenotypes on disease from observational data in the presence of (assumed) unmeasured confounding [**?**]. Causal inference in the form of MR provides a principled foundation for postulating and testing causal effects.

In this work, we propose combining MR techniques with neural network generated data to test for a hypothesized causal relationship between transcription factor binding and chromatin accessibility. In doing so, we hope to enable trustworthy, interpretable estimates of causal effects, which may one day enable discoveries of new causal relationships from in-silico data. We test our method by estimating the postulated effect of CTCF binding on chromatin accessibility.

1.1 Related Work

Many deep learning for genomics papers [6, 13, 7] use variants of a technique called in-silico mutagenesis to try and understand how single nucleotide polymorphisms (SNPs) and other simple

Preprint. Under review.

mutations impact outcomes. Related to this, DeepLIFT [**?**], DeepSHAP [8], and saliency maps [**?**] all allow researchers to interpret how input features (individual nucleotides) impact individual predictions. In-silico mutagenesis experiments have replicated known relationships between genetic variants and binding propensity and chromatin accessibility but lack a principled interpretation in terms of causal inference. DeepLIFT, DeepSHAP, and saliency maps excel at interpreting individual predictions for input/prediction pairs but don't provide a mechanism for pooling different inputs' impacts on predictions.

Didelez and Sheehan [5] gives an overview of Mendelian Randomization and its relationship to instrumental variable methods. Burgess et al. [4] uses (and augments) Mendelian Randomization to estimate the impact of a phenotype on disease in a large population – in this case, body mass index's direct and indirect effect on uric acid levels. Zhao et al. [11] extends Mendelian Randomization to a setting in which we have multiple *approximately valid* instrumental variables, each of which may act weakly on the exposure and outcome. This is directly related to our situation, in which each sequence and mutation for which we predict binding and accessibility serves as its own instrumental variable. Our work therefore heavily relies upon the method described in [11] to test for a causal effect in neural network generated data.

2 Methods

2.1 Mendelian randomization setup

Following [11], our method seeks to estimate the causal effect of a binary-valued exposure X, transcription factor (TF) binding, on a binary-valued outcome Y, chromatin accessibility, by assuming we can use p gene sequences and corresponding SNPs as *approximately valid* instrumental variables (IVs), $Z_1, \ldots, Z_p$. Figure 1 captures the IV assumptions and assumed causal model in a DAG.

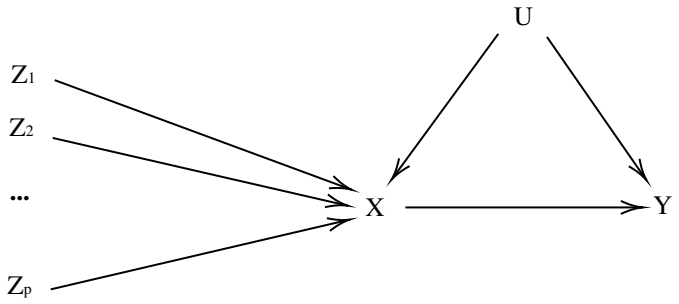

Figure 1: DAG depicting our postulated causal relationships. The structure of the arrows reflects our hypothesis about the causal structure of our phenomenon and also the assumptions required to leverage IV approaches.

As described in [11], we intend to test multiple IV estimation approaches to see how strong assumptions about lack of pleiotropy and IV effect strength affect our results. In particular, we'll compare the causal effect estimate from the Wald estimator with that of the robust adjusted profile score (RAPS). The Wald estimator estimates the causal effect via dividing each regression coefficient from $Z_i \to Y$ by the corresponding regression coefficient from $Z_i \to X$ and aggregating them together using a meta-analysis method. This works well when we can make strong assumptions about lack of pleiotropy and strong IV to exposure effects, but is biased when these assumptions fail. The RAPS estimator was proposed in [11] and allows for weaker assumptions about pleiotropy and weak IV effects. In particular, RAPS pleiotropy is robust to systematic and idiosyncratic (biased across IVs) pleiotropy and weak IV effects. Thus, we expect it to work better in our case where we can't be sure that sequence characteristics that strongly influence CTCF binding don't also influence other binding effects which might impact chromatin accessibility.

2.2 Transcription factor binding and chromatin accessibility prediction

As described above, our causal effect analysis requires estimates of the effects of different mutations on transcription factor binding and on chromatin accessibility probabilities. Unlike in prior MR

studies where this data comes from observational data, in this work, we generate this data using convolutional neural networks (CNN) trained on ChIP-seq and DNase-seq data respectively.

For the TF binding prediction model we will be training a CNN using pytorch on data from the ENCODE database. As for the chromatin accessibility model, we intend to initially use a pre-trained model of Basset. Basset is a tool which allows researchers to train a deep CNN that is able to learn accurate models of DNA sequence activity: accessibility (via DNaseI-seq or ATAC-seq), protein binding (via ChIP-seq), and chromatin state ([6]). Depending on the results we observe, we may ultimately train our own Basset model on different data rather than relying on the pre-trained one. Since Basset's pre-trained model uses human genome hg19 as its reference, we will also be using hg19 as the reference genome for training the TF binding model. We expect both models to achieve 90%+ accuracy, but it is worth discussing the implications of over fitting (see section 3: results and discussion).

We intend to generate IVs by means of in-silico mutagenesis i.e. by mutating the sequences from the validation set (which data set we draw from is yet unclear, see section 3 for discussion). Such mutations will most certainly include SNPs, but we are also examining the possibly of other forms of mutations such as multiple SNPs, insertions, and deletions. Another question which must be addressed is whether or not a single sequence may be used as the basis for multiple IVs, as it is may violate the assumptions required for the inference to work. In order to find strong IVs we plan on performing an in-silico saturated mutagenesis. This will enable us to create a saliency map of which mutations have the greatest impact on TF binding and thus select those specific mutations as our IVs. An IV's strength may be measured using statistical analyses such as ordinary linear regression or generalized method of moments ([3]).

3 Results and discussion

We've currently trained a relatively simple CTCF predictor using a 3-layer CNN architecture. This CNN currently achieves 96% accuracy at predicting CTCF binding on the dataset provided as part of assignment 1, which is drawn from a human lung cancer cell. We intend to eventually replace this dataset with one more compatible with Basset's training and test datasets but first want to test our MR technique end-to-end with data generated from this and the pretrained Basset model. As a pre-requisite to generating our IV data, we're going to ensure that the CTCF predictor's accuracy rises to 98%.

Once we've done this, we'll use saliency maps to identify strong IVs as described above. This will enable us to perform a variety of statistical tests to evaluate our IV's strength and test Wald's method and RAPS with data generated by our two neural networks. From here, our next steps will be determined by the strength of the causal effect we observe. If we observe a strong effect, we'll focus on trying to validate our method.

References

[1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831, 2015.

[2] Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome biology*, 18(1):67, 2017.

[3] C. F. Baum. Instrumental variables and gmm: Estimation and testing. *Stata Journal*, 3(1):1–31(31), 2003. URL `http://www.stata-journal.com/article.html?article=st0030`.

[4] Stephen Burgess, Rhian M Daniel, Adam S Butterworth, Simon G Thompson, and EPIC-InterAct Consortium. Network mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International journal of epidemiology*, 44 (2):484–495, 2014.

[5] Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.*, 16(4):309–330, August 2007.

[6] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

[7] Peter K Koo, Praveen Anand, Steffan B Paul, and Sean R Eddy. Inferring sequence-structure preferences of rna-binding proteins with convolutional residual networks. *bioRxiv*, page 418459, 2018.

[8] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[9] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[10] Jun Zhang, Qingcai Chen, and Bin Liu. Deepdrbp-2l: a new genome annotation predictor for identifying dna binding proteins and rna binding proteins using convolutional neural network and long short-term memory. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[11] Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S. Small. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score, 2018.

[12] Jinfang Zheng, Xiaoli Zhang, Xunyi Zhao, Xiaoxue Tong, Xu Hong, Juan Xie, and Shiyong Liu. Deep-rbppred: Predicting rna binding proteins in the proteome scale based on deep learning. *Scientific reports*, 8(1):15264, 2018.

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