Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Deep Mendelian Randomization: Identifying and Verifying Genomic Deep Learning Models' Causal Knowledge

Stephen Malina SDM2181@COLUMBIA.EDU

David A. Knowles DAK2173@COLUMBIA.EDU

Editor: N/A

1. Outline

Why did I do this work?

1. Establish confidence in our models by validating they learn the qualitative relationships we're confident exist.

2. Identify new potential qualitative relationships to investigate with experiments.

What work relates to ours?

1. Work applying DL models to epigenomic data. We're just building on this.

2. Interpretability work like DeepLIFT analyzing what DL models learn. Tends to focus on individual samples or motifs rather than high-level relationships.

What does our method do? Identify and validate causal relationships learned by genomic DL models.

What do our results mean? Finding that our models seem to learn the causal relationships we expect

What hypotheses did we test? Whether our method could identify the right quantitative and qualitative causal relationships between epigenomic features.

What did we learn?

What did/didn't work? In both simulation and real experiments, Deep MR mostly recovers the order of causal relationships but not the exact magnitude.

What experiments did we do?

Why does is matter? Why should a reader care? Having a method that can verify genomic DL models learn the causal relationships we know exist would increase our confidence in them.

What work would we do next to expand on this project?

Guidance for the reader

Strategy

Things to look out for

- Method's assumptions derived from traditional MR assumptions.

1.1 Related Work

In recent years, many researchers have applied deep learning to achieve impressive results at predicting transcriptomic features such as transcription factor (TF) binding Alipanahi et al. (2015); Zhou and Troyanskaya (2015), chromatin accessibility Kelley et al. (2016); Zhou and Troyanskaya (2015), RNA binding protein (RBP) binding Zheng et al. (2018); Zhang et al. (2019); Koo et al. (2018), and DNA methylation states Angermueller et al. (2017) 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.

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.

Summary

Experiments

- Test method on regression model trained on simulated data with known ground truth and simple generative process - two TFs where one's binding causes the other's.
- Applied method to two jointly trained models, one regression (BPNet) and one classification (DeepSEA), from the literature.

Conclusions

2. Experiments

2.1 Simulation

2.2

3. Discussion

3.1 Challenges

3.1.1 Mendelian Randomization requires strong assumptions

3.1.2 Obtaining calibrated standard error estimates for sequence-level causal effect estimates

3.1.3 Deep MR's accuracy is limited by the accuracy of the underlying model

References

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.

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.

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.

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.

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.

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.

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